SAPIENZA
UNIVERSITÀ DI ROMA

Dottorato di Ricerca in Statistica Metodologica

Tesi di Dottorato XXV Ciclo – 2009/2012

Dipartimento di Scienze Statistiche

New methods for capture-recapture modelling with behavioural response and individual heterogeneity

Danilo Alunni Fegatelli

Supervisor:
Prof. Luca Tardella

This work is composed of three different parts. In Part I it is provided an overview on capture-recapture analysis whose goals, motivations and applications are briefly sketched. It is also proposed a classification of capture-recapture modeling based on five different assumptions characterizing models and data: i) discrete-time data or continuous-time data, ii) closed-population or open-population, iii) sources of variability on (re)capture probabilities, iv) dependence structures among units in the population, v) possibility of misclassification. Moreover, the main methodological issues appearing in capture-recapture analysis are highlighted. The most important is the fact that, differently from the standard-regular statistical settings, the main parameter of interest is discrete, finite although possible unbounded and it affects the range of observable sample space.

Part II and III instead are concerned with two different features of the capture-recapture modeling. More specifically, in the first chapter of Part II, in the context of discrete-time capture-recapture experiments, we will deal with the so called behavioural models. The individual capture history of each unit is represented as a row of a binary matrix and it is conceived as a longitudinal data to allow the modeling of the behavioural effect to capture and hence the dependence structure among the capture occasions. In the literature different model frameworks have been proposed to handle different features of the behavioural effect to capture. As starting point we will review a rather general approach proposed in Farcomeni (2011). This model framework, differently from the most traditional log-linear models, reparametrizes the contingency table probabilities of all possible capture histories in terms of conditional probabilities. In the original paper the conditional maximum likelihood is used as inferential approach. This is in fact the most commonly used approach in most of discrete-time capture-recapture models. However, we highlight that such approach may lead, with positive probability, to inferential pathologies such as unbounded estimates for the finite size of the population. We characterize the occurrence of such likelihood failures within a very general class of behavioural effect models where the probability of never being captured during whole experiment depends on one parameter only. We will show also that the likelihood failure problem is not completely overcome if one uses the unconditional likelihood as alternative approach to make inference on the population size N. Alternatively we propose a fully Bayesian approach pointing out that it completely overcomes the likelihood failure phenomenon. The overall improved performance of alternative Bayesian estimators is investigated under different non-informative prior distributions verifying their comparative merits with both simulated and real data.
In the second chapter of Part II we propose a new alternative model framework based on a meaningful numerical covariate associated, for each binary outcome, to

the previous partial capture history. We show how the use of a suitable ordering and scaling of the progressive partial history of the individual data can be used to model the behavioural effect to capture. The proposed ordering is grounded on the binary representation of integers. We show how appropriate rescalings can be exploited as a suitable quantitative individual time-varying covariate to be embedded in a very general framework such as a generalized logistic model. In fact one can consider the logit of the conditional probability of each binary outcome regressed as a suitable function of the previous partial capture history. A large class of alternative parsimonious sub-models can be explored considering such function as a possibly continuous numerical covariate or grouped as a categorical covariate. We will show how the derivation of the unconditional maximum likelihood estimator can be easily carried out by maximizing the profile likelihood of the population size from the standard output of GLM routines of any statistical software. A similar logistic model structure has been previously sketched in Huggins (1989) and Alho (1990) although the focus there was in developing conditional likelihood estimates in the presence of individual covariates different from partial capture histories. We will show how the classical unconditional approach sometimes can lead to an almost flat likelihood. In order to overcome this issue and get stable inference we propose again a Bayesian analysis. We discuss two alternative approaches which allow to easily implement the Bayesian analysis in this context: a data augmentation approach proposed in Royle et al. (2007) and a more standard approach based on a customized Metropolis-within-Gibbs algorithm. Once again a simulation study will show how the Bayesian analysis still provides improved inference in terms of both point and interval estimates: smaller root mean square error and shorter interval estimates in the presence of a frequentist coverage corresponding to the level of highest posterior density region.

In Part III we deal with flexible statistical models for count data deriving a new tool to infer on an unknown population size in the presence of individual heterogeneity only: no other sources of variability such as behavioural and time effect are considered. In this case we model data corresponding to the number of captures occurred for each unit during the trapping period conveniently summarized in terms of the so called frequencies of frequencies. Differently from the previous behavioural context we can consider also data collected during a continuous-time experiment and data where the number of recaptures has not been a priori bounded. Hence, the total number of captures for each unit can be considered driven by a counting process in a generic time interval representing the trapping period. The summarization of observed data by frequency of frequencies is also used in other scientific problems such as species richness where the goal is to estimate the number of species in a

population based on observed sample count data. Several model frameworks have been adopted to handle data expressed as a series of counts. In most of them the main building block is the Poisson distribution for the individual count data. In order to account for individual heterogeneity of the recapture rate instead of considering an overparameterized model, where a specific individual rate is associated to each unit, we opt for a hierarchical approach using a mixing distribution so that each individual catch rate is a random realization drawn from it. The choice of the mixing distribution obviously influences the model and may restricts its flexibility. In fact an unsuitable choice of the mixing distribution in fact can yield to a systematic distortion of the resulting analyses. In the literature different model settings and inferential approaches have been proposed. As a reference point we will use a recent approach proposed in Wang (2010). In this work it is considered a flexible Poisson compound gamma model estimating the mixture by a penalized non-parametric maximum likelihood approach and then using a least-squares cross-validation procedure for the choice of the common shape parameter. In order to evaluate his approach the author compared his inferential procedure procedure with several classical estimators via simulation and analyses of real data. His approach turns out to be an improvement over many other alternative approaches where the mixing distribution is not restricted to a finite dimensional parametric family. As alternative we will propose a fully Bayesian non-parametric estimate of the population size based reparameterization of the mixture likelihood function in terms of a finite number of moments of a suitable mixing distribution. We compare our nonparametric Bayesian approach implementing a simulation study according to the same setting considered in Wang (2010). Results show that our proposal performs well in terms of point estimates and coverage although slightly more biased than Wang's procedure. Often our proposal yields improved inference: smaller mean square error and frequentists coverage of our Bayesian credible intervals close to the nominal and actual value of the corresponding competitor. The good performances of our approach with simulated data are confirmed by many real data analyses. The resulting estimates are coherent with the underlying scientific knowledge. Moreover, in the examples where the true population size is known in advance our point estimates are close to the truth and the interval estimates always contain the real values.

Contents

Part I

Introduction on Capture-Recapture Analyses

Chapter 1

Capture-Recapture to infer the unknown population size

Capture-recapture methods are statistical tools which allow for the estimation of an unknown population size based on partial observation of this population. The basic idea of capture-recapture analysis is to sample the population several times and then using recapture information, also called overlap information, to estimate the number of uncaptured units in the experiment. Intuitively, when the number of recaptures is low we can infer that the size can be much larger than the number of distinct observed units. On the other hand when the recapture rate is high then we are likely to have captured most of the units in the target population. Obviously, the number of captures and recaptures depends on the population size but it is also affected by the design of the (re)capture plan and by the individual characteristics of the units which correspond to individual capture probabilities and their joint probabilistic structure. In the following we denote by N the true population size and we consider it as the main parameter of interest.

A capture-recapture idea was employed for the first time in 1786 by Laplace to estimate the population size of France. However, capture-recapture methods were formalized roughly a century later in the biological science to estimate the size of a finite wild animal population (Petersen, 1896; Lincoln, 1930). In fact, in that scientific area it is recognized that it is almost impossible to make a census of a wild animal population and hence there is need to estimate the size of the target population through incomplete samples. Ecology has been one of the original fields of development of capture-recapture models and methods although many other fields such as epidemiology, software reliability, genetics, etc. make nowadays extensive use of capture-recapture models and promote further developments.

In epidemiology the purpose of many surveillance studies is to estimate the size of

a diseased (cancer, diabetes, drug use, etc.) population by merging several existing but incomplete lists of names. Regarding each list as a trapping sample and identification numbers and/or names as tags or marks we are in the same situation of a capture-recapture experiment for animal population. The most relevant difference in epidemiological environment lies in the fact that there is a natural time ordering in sampling wild animal population, but generally no such order exists (or is unavailable) when recording units in different lists for an epidemiological study. Moreover, although sometimes such order exists it may vary with individuals.
The number of bugs or faults in a software is an important measure for evaluating software reliability. All the bugs in a software can be considered as the target population. Usually software are debugged independently by several experts. The bugs detected by each expert is seen as a single trapping occasion and hence a capture-recapture model can be applied to estimate the number of undetected bugs.
Other recent fields of application for capture-recapture methods also include the estimation of the number of species in a community and the estimation of the number of expressed genes in genetic applications where the collection of expressed tags represents a random sample of the entire target population.
The Lincoln-Petersen (L-P) estimator can be considered the most basic method in capture-recapture analysis. The L-P estimator is based on 2 sampling operations (capture or trapping occasions) which determine a partial survey of the entire population but it allows to make inference on the population size. In the first occasion a sample of individuals is captured, marked and then released back into the population and then a second sample is observed. The L-P method assumes that the population is closed, all members of the population are equally likely to be marked and recaptured and marked units are randomly distributed in the population at the time of recapture. This means that all units have the same capture probability and this probability does not change if the unit is captured. Furthermore, all units act independently from each others. Notice that there are only four possible encounter histories: captured on occasion 1 and recaptured on occasion 2, captured on occasion 1 and not recaptured on occasion 2, not captured on occasion 1 and captured on occasion 2, not captured at either occasion 1 or occasion 2. Let us denote by (11), (10), (01), (00) these capture histories. Clearly, the number of individuals with history (00) is not observable and hence it has to be estimated. Let z_{11} be the number of individuals captured on both occasions. Analogously, let z_{10} be the number of individuals captured in the first occasion only and z_{01} be the number of units captured on the second occasion only. The number of unobserved units will be z_{00}. Let $n_1 = z_{11} + z_{10}$ be the total number of individuals captured on the first occasion. In the same way we define the total number of individuals captured on the second occasion: $n_2 = z_{11} + z_{01}$. Finally, let $m_2 = z_{11}$ be the number of individuals captured

on both occasions. Under the hypotheses expressed above, the L-P estimator can be intuitively derived from the relation

$$\frac{n_1}{N} \simeq \frac{m_2}{n_2}$$

The rate of units captured in the first occasion is thought to be equivalent to the rate of marked units observed (recaptured) over the total number of units captured in the second occasion. Hence, the L-P estimator is

$$\hat{N}_{LP} = \frac{n_1 \cdot n_2}{m_2}$$

Since the early 30's several contributions to this methodology began to formalize in a more rigorous way the problem by specifying an underlying statistical model and considering more complex sampling strategies for instance with more than two capture occasions (see Amstrup et al. (2005), Chao (2001) for a recent overview on capture-recapture modeling)

1.1 Classification of capture-recapture models

Capture-recapture models can be classified in several ways depending on the available sampling strategies and also on the assumptions and hypotheses adopted. Five principal characteristics of a capture-recapture model can be considered as basis for a sound classification:

- **Discrete-Continuous time**
  In a discrete-time experiment, the target population is sampled over a certain number of capture occasions and, for each occasion, any unit captured can be counted only once. In a continuous-time experiment there is no fixed time where the units are subjected to capture. There is a time-interval where the units are under observation and for each unit is recorded the exact time when it is captured. A continuous-time experiment can be rearranged to make use of a discrete-time model by dividing the time-interval and considering each sub-interval as a capture occasion. On the other hand it is not possible to formalize a discrete-time experiment as a continuous-time model.

- **Closed-Open population**
  In a closed capture-recapture model the unknown population size N is assumed to be constant with no birth-death or immigration-emigration during the all sample stages. In an open capture-recapture model the target population can

vary during the time of the experiment. Obviously models which allow to consider an open population include the closed capture-recapture model as a special case. Moreover, for open population problems the actual population size at a specific fixed time point could not be the main parameter of interest or at least not the only one. In fact, for example, parameters associated to the probability of remaining in the population could be of interest as well as the probability of entering.

- **Sources of variability**
  In the literature there is a classical tripartition of the sources of variability which can affect the probability that a particular unit is caught in one of the trapping occasions: i) behavioural variability due to the change of behaviour of each unit after trapping experience; ii) individual heterogeneity due to observable or unobservable specific characteristics of each unit (sex, age, weight, individual propensity of being captured, etc); iii) temporal, due to the external conditions such as weather, season, trapping effort, etc. which can influence the success of the specific trapping occasion.

- **Dependence of captures among distinct units**
  A usual hypothesis in capture-recapture analyses is that all units act independently. Nevertheless, in some cases, especially with wild populations, units can act in herd leading to a certain dependence structure on the catchability among the units (Fattorini et al. 2007).

- **Misclassification**
  Another typical assumption is that the units do not lose their marks and all tags are recorded correctly. However, this assumption is not always true and hence it is possible to consider misclassification among units in the appropriate model (Link et al. 2010, Tancredi & Liseo 2011).

The most convenient situation, which can be considered as the starting point in capture-recapture analysis, is when the population size is considered to be fixed, no misclassification is allowed and all units act independently. The Lincoln-Petersen estimator is developed under these basic assumptions. Of course it is possible to relax one or more assumptions for example allowing open population and/or misclassification and/or considering a dependence structure among the units to make the model more realistic and closer to the actual conditions underlying the sampling context. However, at the same time, the model becomes much more complex with a higher number of parameters. Following the Occam's razor principle one should never make the model more complex than what it is actually needed. On

the other hand, many times the trapping experiment can be designed so that the basic assumptions are not so far from the reality. For instance, experiment can be planned in a very short time and in a well bounded place so that the population corresponding to this sampling conditions can be considered to be closed and its size constant. Moreover, the tagging process should be as reliable as possible so that the misclassification can be excluded. Finally units often do not act in groups or herds and so it is possible to consider that each unit acts independently. On the other hand when it is not the case one cannot bypass this issue by careful planning.

1.2 Sources of variability and data representation

In this thesis we will adopt the following basic assumptions of the statistical model: we will consider closed population, absence of misclassification and independence among units addressing only which source of variability has to be considered, how such variability can be modelled and whether or not we can deal with a discrete-time or continuous-time model. Those two features of the capture-recapture context influence the complexity of the model and, as we will see, it can also determine how data have to be represented and summarized. From the taxonomy introduced in Otis et al. (1978) we will consider alternative classes of models associated to each source of variability in capture probabilities. We will denote with $\mathcal{M}_B$ the class of models which allow behavioural effect and, analogously, with $\mathcal{M}_T$ and $\mathcal{M}_H$ the classes of models which allow time effect and individual heterogeneity respectively. It is possible to consider more than one source together denoting the corresponding classes of models with $\mathcal{M}_{BT}$, $\mathcal{M}_{BH}$, $\mathcal{M}_{TH}$ and $\mathcal{M}_{BTH}$. When no source is considered we will denote the basic or null model with M_0. In a discrete-time model each source of variability leads to different dependence structure among the samples. Consider t trapping occasions. Data can be (ideally) represented as an $N \times t$ binary matrix $\mathbf{X} = [x_{ij}]$ where

$$x_{ij} = \begin{cases} 1 & \text{unit } i\text{-th is captured at time } j \\ 0 & \text{otherwise} \end{cases}$$

Let us denote with M the number of distinct units observed, and hence captured at least once in the experiment. A typical way to organize the data is to label the observed units from 1 to M and those not captured from $M+1$ to N. Hence, the matrix $\mathbf{X}$ can be seen as follows

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_{obs} \\ \mathbf{X}_{mis} \end{bmatrix}$$

where $\mathbf{X}_{obs}$ is an $M \times t$ binary matrix representing the observed data and $\mathbf{X}_{mis}$ is an $(N - M) \times t$ matrix of unobservable zeros.
Let $p_{ij} = Pr(X_{ij} = 1)$ be the probability that unit i is captured at time j for $i = 1, \ldots, N$ and $j = 1, \ldots, t$. When the null model M_0 is considered the capture occurrence during subsequent occasions are i.i.d. binary vectors so that

$$p_{ij} = p \qquad \forall i = 1, \ldots, N \quad ; \quad \forall j = 1, \ldots, t$$

The capture probability is the same for each unit and this probability does not change from occasion to occasion and hence it does not depend on the previous capture history.
Considering time effect only in the model corresponds to individual captures which are independent but not equally distributed so that the capture probability is the same for each unit but vary from occasion to occasion

$$p_{ij} = p_j \qquad \forall i = 1, \ldots, N \quad ; \quad \forall j = 1, \ldots, t$$

Each occasion has a proper capture probability independently of what has occurred in the other occasions.
When the individual heterogeneity effect is considered the probabilities vary from unit to unit but they do not vary form occasion to occasion

$$p_{ij} = p_i \qquad \forall i = 1, \ldots, N \quad ; \quad \forall j = 1, \ldots, t$$

In this case the capture occurrences in the subsequent occasions can be considered as exchangeable binary outcomes and it does not matter when captures have occurred, but only how many have occurred during whole trapping stages.
When behavioural effect is involved in the model the capture probability at each time depends on the capture status in the previous occasions. In this case one should avoid the previous notation p_{ij}. A very general way to formalize behavioural dependence could be to express the joint probability of binary outcomes in terms of the longitudinal sequence of conditional probabilities denoted with $p_j(x_{i1}, \ldots, x_{ij-1})$ as follows

$$p_j(x_{i1}, ..., x_{ij-1}) = Pr(X_{ij} = 1 | x_{i1}, ..., x_{ij-1}) \quad \forall i = 1, \ldots, N \quad ; \quad \forall j = 1, \ldots, t$$

Units with the same partial capture history in the previous $j - 1$ occasions have the same probability of being captured at time j. In this case, it is natural to consider the capture-recapture experiment as a longitudinal study where each unit is observed several times and at each time the results depend on the previous status and hence there is a longitudinal dependence structure among capture occurrences to be modelled.

If we are in presence of a continuous time model an appropriate way to formalize the observed data is to consider a counting processes for each unit. If the capture-recapture experiment is performed in the interval $[0, \tau]$ one can denote with $C_i(s)$ the number of times the i-th unit is captured in the interval $[0, s]$ for $0 \leq s \leq \tau$ and $i = 1, \ldots, N$. Each $\{C_i(s) : 0 \leq s \leq \tau\}$ is the individual trajectory of a continuous-time counting process which can be parameterized by a capture intensity function $\lambda_i(s)$ where

$$\lambda_i(s) : \lambda_i(s)ds = Pr(dC_i(s) = 1)$$

The notation $dC_i(s) = 1$ can be roughly interpreted as the capture occurrence of unit i in an infinitesimal time interval around time s. The capture intensity $\lambda_i(s)$ plays a similar role of the capture probability p_{ij} in a discrete-time model. Analogously, for continuous time models the sources of variability in capture probabilities lead to similar dependence structure in the capture-recapture counting process. When no sources are considered the capture intensity is constant for whole experiment

$$\lambda_i(s) = \lambda$$

In models where time effect is allowed the intensity is the same for all units but changes by time without considering what has happened previously.

$$\lambda_i(s) = \lambda(s)$$

When an individual heterogeneity effect is considered for each unit it is specified with an individual specific intensity which remains constant for the whole experiment

$$\lambda_i(s) = \lambda_i$$

Considering behavioural effect in the model the capture intensity at each time depends only on the partial capture history occurred

$$\lambda_i(s) \quad \Rightarrow \quad \lambda(s|\mathcal{F}_s^-)ds = Pr(dC_i(s) = 1|\mathcal{F}_s^-)$$

where $\mathcal{F}_s^-$ represents the capture history that has happened up to time t.
Now we point out the fact that the way of collecting data influences and in many cases limits the choice of the model assumed. For example if for each unit is not recorded the exact time of all captures it is not possible to adopt a continuous-time approach. Similarly, if it has been collected only the number of captures occurred for each unit it is not possible to consider a behavioural or time effect in the model because it is not possible to analyse the capture sequences and hence it is not possible to model how the captures are occurred longitudinally over time. As introduced above in a discrete-time context the binary representation formalized

with the matrix $\mathbf{X}$ is the most general way to express the longitudinal structure of capture-recapture data. However, it is important to highlight the fact that the labelling $1, \ldots, N$ is conventional. As we will see in the following sections this aspect of the data is formalized in the likelihood function trough a combinatorial coefficient proportional to $\binom{N}{M}$. Data can be always summarized by the counts of all the entire capture histories observed which are sufficient statistics in any context where there is no individual characteristic which is observed other than the whole individual capture histories. However, they are not always a minimal sufficient statistic. The choice of an appropriate data summarization depends on the sources of variability involved in the model. When no sources of variability is considered the minimal sufficient statistic is the total number of captures (and recaptures) occurred together with the number of distinct units observed during all trapping stages. When time effect only is considered the number of captures occurred on each occasion together with M is the minimal sufficient statistic. When it is considered a model belonging to $\mathcal{M}_B$ the observed data can be summarized by the minimal sufficient statistics represented by the number of times that specific partial capture histories, which define a specific behavioural model, have occurred during the experiment. Finally, when a heterogeneity effect is assumed an alternative (less expensive) way to express the data is to record for each unit only the number of times that the unit is captured in the whole trapping stages instead of the all binary sequence. In fact, data can be summarized into the so called frequencies of frequencies which represent for each observed count the number of units with a particular number of (re)captures. Analogous considerations hold in a continuous-time context. Notice that in the discrete-time experiment the maximum number of possible captures is fixed. In fact each unit can be captured at most t times. On the other hand this is not true in the continuous case where theoretically for each unit the number of recaptures is not necessarily upperbounded.

In the discrete-time case, capture-recapture experiment can be also formalized in the framework of log-linear models (Bishop et al. 1975). In this approach data are represented as an incomplete 2^t contingency table. Each cell in the table represents one of the entire observable capture history and its count corresponds to the number of units with this specific entire capture history. To better understand, we use a three-occasion case ($t = 3$) as an illustrative example. There are eight possible entire capture histories represented as different cells

$$\{(0,0,0),(1,0,0),(0,1,0),(0,0,1),(1,1,0),(1,0,1),(0,1,1),(1,1,1)\}$$

Obviously, the count of the cell corresponding to the unobserved units is structurally missing because the number of units with capture history $(0,0,0)$ is unknown. Let

$\pi_{v_1v_2v_3}$ be defined as follows

$$\pi_{v_1v_2v_3} = Pr(X_{i1} = v_1, X_{i2} = v_2, X_{i3} = v_3) \quad \forall i = 1, \ldots, N$$

where $v_1, v_2, v_3 \in \{0, 1\}$. The log-linear model reparameterizes the probabilities $\pi_{v_1v_2v_3}$ by considering the logarithm of the expected value of each observable cell. In the three-occasion case the most general (saturated) log-linear model can be expressed as follow

$$\begin{aligned} \log\left[E\left(z_{v_1v_2v_3}\right)\right] &= u + u_1 I(v_1 = 1) + u_2 I(v_2 = 1) + u_3 I(v_3 = 1) \\ &+ u_{12} I(v_1 = v_2 = 1) + u_{13} I(v_1 = v_3 = 1) + u_{23} I(v_2 = v_3 = 1) \\ &+ u_{123} I(v_1 = v_2 = v_3 = 1) \end{aligned}$$

where $z_{v_1v_2v_3}$ is the value of the generic cell (v_1, v_2, v_3) for $v_1, v_2, v_3 \in \{0, 1\}$, and $I(A)$ is the indicator function of the event A. Notice that there are seven observed cells, whereas there are eight parameters in the model. Therefore, it is usually assumed that there is no three-occasion interaction term, i.e., $u_{123} = 0$. Models in the classes $\mathcal{M}_T$, $\mathcal{M}_B$ and $\mathcal{M}_H$ can be obtained as special case of log-linear approach.
For example it is easy to verify that independent model with main effect terms only is equivalent to model M_t described in Otis et al. (1978).

1.3 Methodological issues in inferring capture-recapture models

A very simple way to understand intuitively the estimate of the population size N as parameter of interest in capture-recapture models is to consider the following relation

$$N = N(1 - P_0) + NP_0 \approx M + NP_0$$

where P_0 is the probability of recapture never being observed during all capture occasions. The probability of never being captured is usually unknown and hence an estimate is needed. Given the previous relation the Horvitz-Thompson estimator can be derived as follows

$$\hat{N}_{HT} = \frac{M}{1 - \hat{P}_0}$$

Notice also that the parameter P_0 is in fact a function of the unknown model parameters and the number of parameters involved, as well as the function itself, depend

on the model structure. Although one could consider the probability P_0 as a nuisance parameter from the formula of the Horvitz-Thompson estimator it follows that it has a crucial role on the estimation of the population size. Indeed, as we will see in the next sections, this intuitive representation of the estimate of N in a capture-recapture problem corresponds to the formal derivation of the estimate of N by means of two alternative inferential approaches based on the likelihood function: the unconditional likelihood approach and the conditional likelihood approach.
This over-simplified way of regarding estimation in a general capture-recapture problem hides the presence of several inferential complications which are not present in standard parametric models. These complications are related to the following key points:

- **Non-regular statistical model**:
  Capture-recapture models typically violate the conditions configuring a standard regular statistical model. First of all the main parameter of interest is discrete and this prevents one from using derivatives in the estimating procedures. Moreover the support of the observed data is not independent from the population size. In fact, as discussed above, the number of units observed during whole experiment depends on N.

- **Missing data**
  In capture-recapture analysis there are structurally missing data. Notice that the capture histories for uncaptured units are known: only their number are unknown. In fact in a discrete-time model a capture history corresponding to an unobserved unit is made up by a sequence of t zeros corresponding to no capture in each of the t capture occasions. On the other hand in the continuous case each count $C_i(s)$ is equal to zero for each $s \in [0, \tau]$ and $i = M+1, \ldots, N$. However, we cannot complete the data representation because the number of unobserved units is unknown. Moreover, if other covariates are collected on the observed units to help modelling individual heterogeneity they cannot be gathered for the unobserved unit.

- **Asymptotics**
  Consistency properties in regular statistical models are usually studied for eventually divergent number of observed units. In capture-recapture analyses most of the authors study consistency considering the inferential outcome as N diverges. However, in this case N is an unknown parameter which makes the convergence problem more difficult to be conceived and addressed. Another aspect, often neglected in the literature, to evaluate the eventual behaviour

of the inferential outcome is the one related to the accumulation of evidence gathered as the amount of trapping effort increases. Moreover, differently from N, the number t of trapping occasions, or equivalently τ in a continuous-time experiment, is not a parameter and hence can be somehow planned by researchers.

- **Nuisance parameters**
  Usually, especially in discrete-time analyses, the population size is the only parameter of interest. All the other parameters involved in the model can be considered nuisance parameters and hence they may be treated in different way depending on the inferential approach used. Moreover, as seen above in the simplified illustration of the capture-recapture problem the estimate of the population size is closely linked to the estimate of all nuisance parameters defining P_0.

- **Likelihood pathologies**
  One of the most popular approaches in capture-recapture analysis to make inference on N is to factorize the (unconditional) likelihood in two factors: the first factor corresponds to the so-called conditional likelihood, that is the joint probability of observing the recapture histories of the M observed units conditionally on the fact that they are eventually observed within the planned recapture occasions; the second factor is a residual term also called residual likelihood. Hence the estimate of N is broken down in two steps: one focuses first on the maximization of the so-called conditional likelihood (corresponding to the observed units) deriving an estimation for all model parameters but N while in the second step one maximizes the residual likelihood as a function of N only, plugging in the estimates of all the other nuisance parameters involved in the first factor. Unfortunately, in some cases the conditional likelihood approach could lead to inferential pathologies such as non identifiability problems Link (2003) or likelihood failures ($\hat{N} = \infty$) as we will see in Chapter 2.

All these inferential issues make inference in capture-recapture models very hard even in cases where the model structure is relatively simple.

Part II

Behavioural Effect Modeling

Chapter 2

Likelihood Failure and Improved Inference on Behavioural Capture-Recapture

In the context of discrete-time closed capture-recapture modeling for estimating the unknown size of a finite population it is often required a flexible framework for dealing with a behavioural response to trapping. This will lead us to restrict the attention from the most general embedding model framework denoted with $\mathcal{M}_{TBH}$ in Otis et al. (1978) to the more restrictive $\mathcal{M}_{TB}$ framework. Many alternative settings have been proposed in the literature to account for the variation of capture probability at each occasion depending on the previous capture history. There is a lot of very recent and less recent papers which are concerned with modeling and inferring behavioural patterns. Different approaches have been used ranging from most frequent and classical conditional-likelihood-based inference of Huggins (1989, 1991) to more recent Markov-chain (Yang & Chao 2005) and extensions thereof (Farcomeni 2011), latent class models (Bartolucci & Pennoni 2007), semiparametric covariate dependent approach (Hwang & Huggins 2011) and others (Ramsey & Severns 2010). Inference is typically carried out relying on the so-called conditional likelihood approach. Fewer authors have adopted a Bayesian approach for the simplest permanent behavioural settings (Lee & Chen 1998, Lee et al. 2003, Ghosh & Norris 2005) while alternative estimating approaches have been more recently proposed to cope with behavioural modeling in continuous-time recapture settings (Chaiyapong & Lloyd 1997, Yip et al. 2000, Chao et al. 2000, Hwang et al. 2002). We highlight that the CML approach and also the unconditional maximum likelihood approach (UML) ,based on the maximization of the profile likelihood, may, with positive probability, lead to inferential pathologies such as unbounded estimates for

the finite size of the population. Such annoying phenomenon called *likelihood failure* which also implies some degree of non robustness of the estimator even when it is guaranteed to yield a finite estimate. The occurrence of such likelihood failures is characterized within a very general class of behavioural effect models. This form of degeneracy is not true in general. It is true only sometimes, and especially if you estimate the unknown population size N by means of the maximization of the conditional or profile likelihood. We connect this phenomenon to a problem pointed out similarly by Seber & Whale (1970) in modeling removal studies and later on faced by Carle & Strub (1978) who suggested a weighted likelihood approach as a possible overcome. We also highlight the generality of this likelihood failure phenomenon providing general conditions for its occurrence. We will show that a fully Bayesian approach theoretically overcomes the possible unboundedness of estimates and we propose alternative Bayesian estimators built under different non-informative prior distributions for further investigation. We will compare the conditional and unconditional maximum likelihood estimator (CMLE and UMLE) and the proposed alternative Bayesian estimators via simulation studies providing empirical evidence of overall improved performance of Bayesian alternatives. We also evaluate the performance in a real data application considering the Great-Copper data set.

2.1 Capture-Recapture behavioural effect modeling

Let us consider a discrete-time closed capture-recapture experiment in which the unknown population size N is assumed to be constant and individual trappings are recorded in t consecutive times. Moreover we suppose that all units act independently and there is no misclassification i.e. all individuals are always recorded correctly and do not lose their marks. and that units captured during the study are labelled from 1 to M and those not captured from $M+1$ to N. It is clear that we can observe only the firsts M rows of the matrix $\mathbf{X}$. Denoting with $\mathcal{X} = \{0, 1\}$, the space of all possible capture histories for each unit is $\mathcal{X}^t = \{0, 1\}^t$ while the set of all observable capture histories is $\mathcal{X}_*^t = \mathcal{X}^t \setminus (0, \ldots, 0)$ since the unobserved units are not sampled.

In this work we review the main aspects of modeling the behavioural effect to capture revisiting some general model frameworks proposed in literature. Indeed, mice, voles and small mammals often modify their behaviour after being trapped and this change can reduce or increase the probability of later recaptures. Originally Otis et al. (1978) introduced the basic behavioural model M_b, where individual capture

probabilities vary only once when first capture occurs. Model M_b is the simplest way to consider behavioural effects. In particular it considers an *enduring* effect to capture since the behaviour, and, consequently, the recapture probability change permanently until the end of the experiment. In model M_b the initial capture probability is denoted with p. It is the same for each unit and remains constant from occasion to occasion until the first capture. Once the unit is captured for the first time the (re)capture probability p changes in r and it remains the same until the end of trapping stages. Formally, in order to distinguish the first capture probability from the recapture probability we will make use of the conditioning with respect to the quantity $\sum_{l=1}^{j-1} x_{il}$ corresponding to the number of recaptures prior to the current time j

$$M_b: \begin{cases} Pr(x_{ij}=1 \mid \sum_{l=1}^{j-1} x_{il}=0)=p & \forall i=1,\dots N \ \forall j=1,\dots t \\ Pr(x_{ij}=1 \mid \sum_{l=1}^{j-1} x_{il}>0)=r & \forall i=1,\dots N \ \forall j=2,\dots t \end{cases}$$

where if the upperbound of the summation index is such that $j-1 \leq 0$ then the conditioning event $\sum_{l=1}^{j-1} x_{il} = 0$ is dropped. When $r < p$ the capture probability decreases for all subsequent recaptures and this corresponds to modeling the so called *trap shyness*. This behavioural pattern could be due to the traumatic event associated to the capture experience. On the other hand, when $r > p$ there is the so called *trap happiness* effect.

Alternative model frameworks have been recently proposed to model more flexibly behavioural patterns during trapping stages. Yang & Chao (2005) propose to model the capture history sequence by a bivariate Markov chain in which the states incorporate the information on both capture status (captured/non-captured) and marking status (marked/non-marked). Notice that, obviously, if a unit is captured in the previous occasions it is also marked. Yang-Chao's model allows to handle both enduring effects where individuals exhibit a long lasting behavioural response to capture and the so called *ephemeral* effect where individuals have a short term memory and the capture probabilities depend only on the capture occurrence in the previous occasion. When the marking status is not considered we have the simple first-order Markov chain model allowing for ephemeral effect only. A generalized $k-$th order Markov chain model is considered in Farcomeni (2011) and it is denoted by M_{c_k}. In model M_{c_k}, for each unit, capture probability at some stage j depends only on the capture status of the unit in the previous k occasions. More formally for $k=1$ in model M_{c_1} we have

$$M_{c_1}: \begin{cases} p(x_{ij}=1|x_{ij-1}=0)=p_{(0)}, & \forall i=1,\dots N \ \forall j=1,\dots t \\ p(x_{ij}=1|x_{ij-1}=1)=p_{(1)}, & \forall i=1,\dots N \ \forall j=2,\dots t \end{cases}$$

while for $k = 2$ in model M_{c_2} we have

$$M_{c_2} : \begin{cases} Pr(x_{ij} = 1 | x_{ij-2} = 0, x_{ij-1} = 0) = p_{(00)}, & \forall i = 1, \dots N \ \forall j = 1, \dots t \\ Pr(x_{ij} = 1 | x_{ij-2} = 0, x_{ij-1} = 1) = p_{(01)}, & \forall i = 1, \dots N \ \forall j = 2, \dots t \\ Pr(x_{ij} = 1 | x_{ij-2} = 1, x_{ij-1} = 0) = p_{(10)}, & \forall i = 1, \dots N \ \forall j = 3, \dots t \\ Pr(x_{ij} = 1 | x_{ij-2} = 1, x_{ij-1} = 1) = p_{(11)}, & \forall i = 1, \dots N \ \forall j = 3, \dots t \end{cases}$$

For $k = 1, 2$ if $j - k \leq 0$ the conditioning events related to x_{ij-k} are dropped. We remark that in all the models considered so far the probability of never being observed during all t occasions, denoted by P_0, depends only on one parameter. More precisely we have for the previous models

$$\begin{aligned} M_b &: \quad P_0 = (1 - p)^t \\ M_{c_1} &: \quad P_0 = (1 - p_{(0)})^t \\ M_{c_2} &: \quad P_0 = (1 - p_{(00)})^t \end{aligned}$$

As we will see the probability P_0 plays a crucial role in determining the estimate of the population size.

It is also possible to consider an encompassing model which allows for both ephemeral and enduring effects together and it will be denoted with $M_{c_k b}$. It basically consists of a generalized k-th order Markov chain model where, in correspondence of the same conditioning k-th order event $x_{j-k} = 0, \dots, x_{j-1} = 0$, one distinguishes those histories where a previous first capture has occurred. Only for the partial capture histories formed by k zeroes in the last k occasions we need to specify if a unit is marked or not. In conceiving an appropriate notation for the different capture probabilities the fact that a unit has been captured previously (and hence marked) can be denoted by the digit 0 or 1 before the comma. For example in model $M_{c_3 b}$, $p_{0,(000)}$ is the probability that a unit is captured at a generic stage j given it is not captured previously and hence it is unmarked; while, $p_{1,(000)}$ is the probability that a unit is captured at time j given it is not captured in the previous $k = 3$ stages but it is captured at least once previously and hence it is marked. Indeed, Yang-Chao's model framework corresponds to $M_{c_1 b}$. To better understand let us consider the following capture history

$$(0, 0, 1, 0, 0, 1, 1, 0, 0, 1)$$

for all models described above the associated sequences of the capture-recapture probabilities are

- Model M_b

  $1 - p \quad 1 - p \quad p \quad 1 - r \quad 1 - r \quad r \quad r \quad 1 - r \quad 1 - r \quad r$

- Model M_{c_1}
  1-p_0 1-p_0 p_0 1-p_1 1-p_0 p_0 p_1 1-p_1 1-p_0 1-p_0
- Model M_{c_2}
  1-p_{00} 1-p_{00} p_{00} 1-p_{01} 1-p_{10} p_{00} p_{01} 1-p_{11} 1-p_{10} p_{00}
- Model $M_{c_1 b}$
  1-$p_{0,0}$ 1-$p_{0,0}$ $p_{0,0}$ 1-$p_{1,1}$ 1-$p_{1,0}$ $p_{1,0}$ $p_{1,1}$ 1-$p_{1,1}$ 1-$p_{1,0}$ 1-$p_{1,0}$
- Model $M_{c_2 b}$
  1-$p_{0,00}$ 1-$p_{0,00}$ $p_{0,00}$ 1-$p_{1,01}$ 1-$p_{1,10}$ $p_{1,00}$ $p_{1,01}$ 1-$p_{1,11}$ 1-$p_{1,10}$ $p_{1,00}$

Farcomeni (2011) provides a much more flexible framework based on the capture probabilities conditioned on each possible partial capture history as follows

$$\begin{cases} p_1() = Pr(x_{i1} = 1) \\ p_j(x_{i1}, ..., x_{ij-1}) = Pr(x_{ij} = 1 | x_{i1}, ..., x_{ij-1}) \qquad \forall j > 1 \ , \ \forall (x_{i1}, \ldots, x_{ij-1}) \in \mathcal{X}^{j-1} \end{cases}$$

All these conditional probabilities can be arranged with a natural order in a $2^t - 1$ dimensional vector as follows

$$\mathbf{p} = (p_1(), p_2(0), p_2(1), p_3(0,0), p_3(0,1), p_3(1,0), ..., p_t(0, ..., 0), ..., p_t(1, ..., 1))$$

where, for example, the element $p_3(0,1)$ represents the probability of being captured at time 3 given that the unit is not captured in the first occasion while it is captured in the second occasion. The initial empty brackets () is understood as the absence of previous capture history at time 1. The vector $\mathbf{p}$ can be seen as a convenient reparameterization of the joint probabilities corresponding to all $2^t - 1$ complete capture history configurations in $\mathcal{X}_*^t$. The conditional probabilities, rather than the joint probabilities, are more easily interpreted in the process of modeling the consequences determined by the change of behaviour due to a particular previous trapping history.
Notice that under the saturated reparameterization the probability of never being observed during trapping stages is

$$P_0 = \left[(1 - p_1()) \prod_{j=2}^{t} (1 - p_j(0, \ldots, 0)) \right] \tag{2.1}$$

From the saturated parametrization one can specify a parsimonious nested model based on a suitable partition of the conditional probabilities in $\mathbf{p}$ in terms of equivalence classes. Let H be the set of all partial capture histories: $H = \{$ () , (0), (1),

(00), (10), (01), (11), ...} $= \cup_{j=0}^{t-1} \mathcal{X}^j$ where $\mathcal{X}^0 = \{()\}$. Denote by $\mathcal{H}_B$ one of the possible partitions of H in B disjoint subsets

$$\mathcal{H}_B = \{H_1, \ldots, H_b, \ldots, H_B\}$$

where each $H_b \subset H$. The role of the index set H is to list all the partial capture histories which may yield possible changes in the conditional capture probability depending on the past.
There is a corresponding parameter vector of probabilities denoted with $\mathbf{p}_{\mathcal{H}_B} = (p_{H_1}, \ldots, p_{H_B})$. Define a generic partial capture history $\mathbf{h}$ as follows

$$\mathbf{h} = (h_1, \ldots, h_{l_h}) \tag{2.2}$$

where l_h is the length of the binary vector. The partition of capture histories in equivalence classes is such that

$$\forall \quad \mathbf{h}, \mathbf{h}^{'} \in H_b \quad \Rightarrow \quad p_{(l_h+1)}(\mathbf{h}) = p_{(l_{h'}+1)}(\mathbf{h}^{'}) = p_{H_b} \qquad \forall b = 1, \ldots, B$$

Notice that when there is absence of previous capture history ($\mathbf{h} = ()$) we have $l_h = 0$.
With the partition $\mathcal{H}_B$ of subsets of H representing equivalence classes we make more explicit the fact that the set of very specific constraints formalized in Farcomeni (2011) as $\mathbf{Cp} = \mathbf{0}$ are nothing but a way to identify blocks of conditional probabilities corresponding to the same common value hence reducing the number of free parameters with respect to the saturated model. Indeed in the $\mathbf{Cp} = \mathbf{0}$ formalization the entries of the constraint matrix C must obey further restrictions (only entries -1,0 or 1 and no more that one 1 entry in each column) and this, we believe, is not very natural. No other specific use of those linear constraints are suggested in that paper.
In the following we will denote by $\mathcal{M}$ the class of models based on conditional probabilities parameterization and specified in terms of a suitable partition $\mathcal{H}_B$.
As an example of such formalization based on partitions of subsets of H one can consider a model which assumes that only after being captured for more than 2 times in a row the behaviour of an animal/unit can be affected so that the probability of being trapped again could be lower (or greater). This simple model denoted with $M_{\bullet\bullet}$ can be formalized using the following (bi)partition of the partial capture histories $\mathcal{H}_2(M_{\bullet\bullet}) = \{H_1, H_2\}$ where

$$\begin{cases} H_1 = \{\mathbf{h} \in H : l_h \ < 2\} \cup \{\mathbf{h} \in H : l_h \geq 2 \ , \ h_{l_h-1} + h_{l_h} < 2\} \\ H_2 = H \setminus H_1 \end{cases}$$

As another example, we can build up a model, denoted with $M_{\#}$ where the number of captures occurred may influence the capture probability. The corresponding

partition denoted with $\mathcal{H}_t(M_\#)$ splits the set H in t equivalence classes each corresponding to a specific total number of captures as follows

$$\begin{cases} H_1 = \mathcal{X}^0 \cup \left\{ h \in \cup_{s=1}^{t-1} \mathcal{X}^s : \sum_{s=1}^{l_h} h_s = 0 \right\} \\ H_2 = \left\{ h \in \cup_{s=1}^{t-1} \mathcal{X}^s : \sum_{s=1}^{l_h} h_s = 1 \right\} \\ \dots \\ H_r = \left\{ h \in \cup_{s=1}^{t-1} \mathcal{X}^s : \sum_{s=1}^{l_h} h_s = r-1 \right\} \\ \dots \\ H_{t-1} = \left\{ h \in \cup_{s=1}^{t-1} \mathcal{X}^s : \sum_{s=1}^{l_h} h_s = t-2 \right\} \\ H_t = \left\{ h \in \cup_{s=1}^{t-1} \mathcal{X}^s : \sum_{s=1}^{l_h} h_s = t-1 \right\} \end{cases}$$

Indeed Farcomeni (2011) provides this general framework where the generic partition is rather specified equivalently in terms of linear constraints on $\mathbf{p}$. This constraints are specified by a $2^t - 1 \times 2^t - 1$ matrix $\mathbf{C}$ as follows

$$\mathbf{Cp} = \mathbf{0}$$

where the generic element of the matrix $\mathbf{C}$ denoted by c_{ij} is such that $c_{ij} \in \{0, 1, -1\}$ with the restriction that each column of $\mathbf{C}$ can not have positive and negative values at the same time. The number of free parameters in the constrained model is the number of columns without negative values which are in one-to-one correspondence with the representative elements of each equivalence class. For example, model M_b can be obtained by using two blocks of equality constraints

$$\begin{cases} p_1() = p_2(0) = p_3(0,0) = \dots = p_t(0, \dots, 0) = p \\ p_2(1) = p_3(10) = p_3(01) = \dots = p_t(1, \dots, 1) = r \end{cases}$$

Equivalently model M_b corresponds to a bipartition $\mathcal{H}_2(M_b) = \{H_1, H_2\}$ such that

$$\begin{cases} H_1 = \{(), (0), (00), \dots, (0 \dots 0)\} = \mathcal{X}^0 \cup \left\{ \mathbf{h} \in \cup_{j=1}^{t-1} \mathcal{X}^j : \sum_{j=1}^{l_h} h_j = 0 \right\} \\ H_2 = H \setminus H_1 \end{cases}$$

In the original paper it is also shown that many models proposed in the literature such as model M_0, M_b, M_{c_k}, $M_{c_k b}$, M_t can be recovered as special cases of model with saturated parameterization $\mathbf{p}$ subject to specific linear constraints corresponding to $\mathbf{C}$.

In the following we prefer to index parameters with the partition notation and we refer to the reduced parametrization $\mathbf{p}_{\mathcal{H}_B} = (p_{H_1}, \dots, p_{H_B})$ corresponding to the uniquely identified conditional probabilities associated to the partition $\mathcal{H}_B$.

2.2 Conditional Likelihood Approach and Likelihood Failure

Under individual independence assumption, the likelihood function can be written in terms of saturated parameterization as follows

$$L(N,\mathbf{p}) = \prod_{i=1}^{N} p_1()^{x_{i1}}(1-p_1())^{1-x_{i1}} \prod_{j=2}^{t} p_j(x_{i1},...,x_{ij-1})^{x_{ij}}(1-p_j(x_{i1},...,x_{ij-1}))^{1-x_{ij}}$$

In order to highlight the generality of some pathological likelihood features of behavioural models we focus on the subclass of all models associated to a generic partition $\mathcal{H}_B$ where all the conditioning partial capture histories corresponding to no capture belong to the same partition set, say H_1. This means that all the conditional probabilities $(p_1(), p_2(0), \ldots, p_t(0,\ldots,0))$ determining P_0 as in (2.1) correspond to the same parameter value. Notice that in the class of models we are considering the first partition set denoted as H_1 can contain also other partial capture histories beside those corresponding to no capture. In the following we will denote by $\tilde{\mathcal{M}}$ this special class of models where, by convention, the first set H_1 listed in the partition $\mathcal{H}_B$ contains (at least) all the aforementioned capture histories defining P_0. Of course $\tilde{\mathcal{M}} \subset \mathcal{M}$. It is easy to verify that model M_0, M_b, M_{c_k}, $M_{c_k b}$ belong to $\tilde{\mathcal{M}}$. As an example, we consider model M_{c_1} corresponds to the partition $\mathcal{H}_2(M_{c_1}) = \{H_1, H_2\}$ such that

$$\begin{cases} H_1 = \{(), (0), (00), (10), \ldots\} = \mathcal{X}^0 \cup \left\{\mathbf{h} \in \cup_{j=1}^{t-1}\mathcal{X}^j : h_{l_h} = 0\right\} \\ H_2 = H \setminus H_1 \end{cases}$$

where H_1 contains the void capture history () and all partial capture histories $h = (h_1, \ldots, h_{l_h})$ such that the terminal digit $h_{l_h} = 0$ for $l_h = 1, 2, ..., t-1$. Of course the conditioning capture histories corresponding to no capture are contained in H_1. Analogously, for model M_{c_2} the partition $\mathcal{H}_4(M_{c_2}) = \{H_1, H_2, H_3, H_4\}$ will be

$$\begin{cases} H_1 = \{(), (0), (00), (000), \ldots\} = \mathcal{X}^0 \cup (0) \cup \left\{\mathbf{h} \in \cup_{j=2}^{t-1}\mathcal{X}^j : h_{l_h-1} = 0\ ,\ h_{l_h} = 0\right\} \\ H_2 = \{(1), (01), (001), (101), \ldots\} = (1) \cup \left\{\mathbf{h} \in \cup_{j=2}^{t-1}\mathcal{X}^j : h_{l_h-1} = 1\ ,\ h_{l_h} = 0\right\} \\ H_3 = \{(10), (010), (110), \ldots\} = \left\{\mathbf{h} \in \cup_{j=2}^{t-1}\mathcal{X}^j : h_{l_h-1} = 0\ ,\ h_{l_h} = 1\right\} \\ H_4 = H \setminus (H_1 \cup H_2 \cup H_3) \end{cases}$$

On the other hand, model M_t does not belong to $\tilde{\mathcal{M}}$. It can be expressed as $\mathcal{H}_t(M_t) = \{H_1, \ldots, H_t\}$ where $H_j = \mathcal{X}^{j-1}$ for $j = 1, \ldots, t$. To better understand we consider a discrete capture-recapture experiment with five capture occasions. When

$t = 5$ we have that models M_b, M_{c_1}, M_{c_2} and M_t yield the following partitions

$$\mathcal{H}_2(M_b) \ : \begin{cases} H_1 = \{(), (0), (00), (000), (0000)\} \\ H_2 = \{(1), (10), (01), (11), (100), (010), (001), (110), (101), (011), (111), \\ \qquad (1000), (0100), (0010), (0001)(1100), (1010), (0110), (1110), \\ \qquad (1001), (0101), (0011), (1101), (1011), (0111), (1111)\} \end{cases}$$

$$\mathcal{H}_2(M_{c_1}) : \begin{cases} H_1 = \{(), (0), (00), (10), (000), (100), (010), (110), \\ \qquad (0000), (0100), (0010), (1000), (0110), (1100), (1010), (1110)\} \\ H_2 = \{(1), (01), (11), (001), (101), (011), (111), \\ \qquad (0001), (0011), (0101), (0111), (1001), (1011), (1101)(1111)\} \end{cases}$$

$$\mathcal{H}_4(M_{c_2}) : \begin{cases} H_1 = \{(), (0), (00), (000), (100), (0000), (0100), (1000), (1100)\} \\ H_2 = \{(10), (010), (110), (0010), (0110), (1010), (1110)\} \\ H_3 = \{(1), (01), (001), (101), (0001), (0101), (1001), (1101)\} \\ H_4 = \{(11), (011), (111), (0011), (0111), (1011), (1111)\} \end{cases}$$

$$\mathcal{H}_5(M_t) \ : \begin{cases} H_1 = \{()\} \\ H_2 = \{(0), (1)\} \\ H_3 = \{(00), (01), (10), (11)\} \\ H_4 = \{(000), (001), (010), (011), (100), (101), (110), (111)\} \\ H_5 = \{(0000), (0001), (0010), (0011), (0100), (0101), (0110), (0111) \\ \qquad (1000), (1001), (1010), (1011), (1100), (1101), (1110), (1111)\}\} \end{cases}$$

Notice also that within the class $\tilde{\mathcal{M}}$ all the models such as M_b and $M_{c_k b}$ do have the first set of the partition H_1 containing all and solely the partial capture histories with no capture i.e. with no 1 digit, while models such as M_{c_k} do have H_1 containing also other partial capture histories. Hence, for all models belonging to $\tilde{\mathcal{M}}$ it will be

$$P_0 = (1 - p_{H_1})^t$$

The likelihood function corresponding to the generic model $M_{\mathcal{H}_B} \in \tilde{\mathcal{M}}$ parametrized with the vector of conditional probabilities $\mathbf{p}_{\mathcal{H}_B}$ will have the following form

$$L(N, \mathbf{p}_{\mathcal{H}_B}) \propto \left[\binom{N}{M} p_{H_1}^{n_{(H_1 1)}} (1 - p_{H_1})^{n_{(H_1 0)} + t(N-M)} \right] \prod_{b=2}^{B} p_{H_b}^{n_{(H_b 1)}} (1 - p_{H_b})^{n_{(H_b 0)}} \quad (2.3)$$

where $n_{(H_b0)}$ is the number of times that all the observed units which experience partial capture history $\mathbf{h}$ belonging to H_b are not captured at time $l_h + 1$; similarly $n_{(H_b1)}$ is the number of times that the observed units which experience partial capture history $\mathbf{h}$ belonging to H_b are captured at time $l_h + 1$. Formally $\forall\, b = 1, \ldots, B$

$$n_{(H_b0)} = \sum_{i=1}^{M} \sum_{\mathbf{h} \in H_b} I\left[(x_{i1}, \ldots, x_{il_h}) = \mathbf{h}\ ,\ x_{i(l_h+1)} = 0\right]$$
$$n_{(H_b1)} = \sum_{i=1}^{M} \sum_{\mathbf{h} \in H_b} I\left[(x_{i1}, \ldots, x_{il_h}) = \mathbf{h}\ ,\ x_{i(l_h+1)} = 1\right]$$

These are easily recognized as the sufficient statistics in this model framework. The classical estimation procedure considered in Farcomeni (2011) is based on the factorization of the likelihood function in 2.3 as in Sanathanan (1972) as follows

$$\begin{aligned} L(N, \mathbf{p}_{\mathcal{H}_B}) &\propto \binom{N}{M}(1-P_0)^M P_0^{(N-M)} \times \frac{1}{(1-P_0)^M} \prod_{b=1}^{B} p_{H_b}^{n_{(H_b1)}} (1-p_{H_b})^{n_{(H_b0)}} \\ &= L^r(N, p_{H_1}) \times L^c(\mathbf{p}_{\mathcal{H}_B}) \end{aligned}$$

where L^c is the conditional likelihood while L^r is the residual (binomial) likelihood. The conditional maximum likelihood estimator $\hat{N}_{CMLE}$ of N is obtained in 2 steps: first we compute $\hat{\mathbf{p}}_{\mathcal{H}_B}$ maximizing $L^c(\mathbf{p}_{\mathcal{H}_B})$ and then using $\hat{p}_{H_1} \in \hat{\mathbf{p}}_{\mathcal{H}_B}$ maximize $L^r(N, \hat{p}_{H_1})$ with respect to N. Let $q_{H_1} = 1 - p_{H_1}$; the CMLE of N is given by

$$\hat{N}_{CMLE} = \frac{M}{1 - \hat{q}_{H_1}^t} = \frac{M}{1 - \hat{P}_0} \tag{2.4}$$

where $\hat{q}_{H_1} = 1 - \hat{p}_{H_1}$ must satisfy the conditional likelihood equation

$$\frac{q_{H_1}}{1 - q_{H_1}} \frac{n_{(H_1 1)}}{M} - \frac{t q_{H_1}^t}{1 - q_{H_1}^t} = \frac{n_{(H_1 0)}}{M} (\equiv R_{H_1}) \tag{2.5}$$

Equation (2.5) can be numerically solved and then the estimate $\hat{q}_{H_1} = 1 - \hat{p}_{H_1}$ is plugged into (2.4). This corresponds to the Horvitz-Thompson estimator which can be also derived as the classical maximum likelihood estimator of the number of trials in a binomial experiment when the probability of success is known and it is equal to $1 - \hat{P}_0$.

However in Seber & Whale (1970) it is pointed out for the first time that in a related removal model the conditional likelihood approach may end up with an unbounded estimate $\hat{N}_{CMLE}$ yielding an annoying inferential pathology called likelihood failure. In the removal model of Seber & Whale (1970), similarly to our models in the class $\tilde{\mathcal{M}}$, units act independently and at each trapping time the capture probability is p and it is the same for each unit. When a unit is captured for the first time it is removed from the population. The likelihood function for the removal model is

$$L_R(N, p) = \tbinom{N}{M} p^M (1-p)^{n_{0p} + t(N-M)}$$

where n_{0p} is the number of times that observed units are not captured i.e. $n_{0p} = \sum_{i=1}^{M}\sum_{j=1}^{t} I(\sum_{l=1}^{j} x_{il} = 0)$. Notice that the likelihood for a removal model has the same functional form of the factor within brackets in (2) on the right hand side. Since the CML estimation of N and p_{H_1} from (2) depend only on the expression within brackets it could end up with the same pathological unbounded estimates as the removal model.

Notice that the argument which shows that the estimates of N depend only on the expression within brackets makes all models $M_{\mathcal{H}_B} \in \tilde{\mathcal{M}}$ sharing the same element $H_1 \in \mathcal{H}_B$ equivalent in terms of the resulting estimates of N. For instance the partitions corresponding to models M_b and $M_{c_k b}$ do share the same H_1.

Hence we claim that it is important to be aware of the possible occurrence of likelihood failures within general frameworks for behavioural modeling like the one proposed in Farcomeni (2011) once the conditional maximum likelihood is pursued. In particular we show that it is possible to characterize the likelihood failure occurrence for the generic subclass of models $\tilde{\mathcal{M}}$. Adapting from Seber & Whale (1970) we provide the conditions which guarantee the finiteness and the uniqueness of the CML solution in that class of models.

In order to understand the behaviour of the solving equations (2.4) and (2.5) consider the left-hand side of (2.5) as a function f of q_{H_1}

$$f(q_{H_1}) = \frac{q_{H_1}}{1-q_{H_1}}\frac{n_{(H_1 1)}}{M} - \frac{t q_{H_1}^t}{1-q_{H_1}^t}$$

Notice that we always have $n_{(H_1 1)} \geq M$. In fact, the number of times that observed units with partial capture history $h \in H_1$ are not captured at time $l_h + 1$ is at least M. For models such as M_b, $M_{c_k b}$ and $M_{\#}$ the statistic $n_{(H_1 1)}$ is always equal to M. For $0 \leq q_{H_1} < 1$ we have that

$$\frac{df(q_{H_1})}{dq_{H_1}} = \frac{1}{(1-q_{H_1})^2}\left[1 - \frac{t^2 q_{H_1}^{t-1}(1-q_{H_1})^2}{(1-q_{H_1}^t)^2}\right] > 0$$

hence, $f(q_{H_1})$ is an increasing function in $[0, 1)$. Consider the limit of $f(q_{H_1})$ for $q_{H_1} \to 1^-$; we have to distinguish 2 cases

$$\begin{cases} \lim_{q_{H_1}\to 1^-} f(q_{H_1}) = \frac{1}{2}(t-1) & n_{(H_1 1)} = M \\ \lim_{q_{H_1}\to 1^-} f(q_{H_1}) = \infty & n_{(H_1 1)} > M \end{cases}$$

When $n_{(H_1 1)} = M$ there exists a unique solution $0 < q_{H_1} < 1$ if and only if R_{H_1} defined in (2.5) is such that

$$0 < R_{H_1} < \frac{1}{2}(t-1) \tag{2.6}$$

In fact, $R_{H_1} > (t-1)/2$ implies that q_{H_1} maximizing the conditional likelihood will be a boundary estimate $\hat{q}_{H_1} = 1$ which implies $\hat{P}_0 = 1$ and hence an infinite estimate of the population size $\hat{N}_{CMLE} = M/(1-\hat{P}_0) = \infty$ (likelihood failure!). Of course restricting q_{H_1} in $(0,1)$ does not overcome this issue. On the other hand, when $n_{(H_1 1)} > M$ the fact that $\lim_{q_{H_1}\to 1^-} f(q_{H_1}) = \infty$ leads to a unique solution $0 < q_{H_1} < 1$ and hence a finite estimate $\hat{N}_{CMLE}$.
In order to evaluate the likelihood failure occurrence we compute the expected value and the variance of $R_{H_1} = n_{(H_1 0)}/M$ denoted with $E(R_{H_1})$ and $V(R_{H_1})$ respectively. Consider enduring effects to capture such as those in models M_b, $M_{c_k b}$, etc. In these models only is relatively simple to obtain $E(R_{H_1})$ and $V(R_{H_1})$ in closed form. In fact, only for models which allow an enduring effect we have that M is a binomial random variable with parameters N and $1 - P_0 = 1 - (1-p_{H_1})^t$ while $n_{(H_1 0)}$ given $M > 0$ is a sum of M truncated geometric random variables with parameters p_{H_1} and t and truncated support $\{0, 1, \ldots, t-1\}$. Hence, using the standard formulas for the expected value and variance we have

$$\begin{aligned} E(R_{H_1}) &= E\left[E\left(R_{H_1}|M>0\right)\right] = E\left[\frac{E\left(n_{(H_1 0)}|M>0\right)}{M}\right] = \\ & E\left[\frac{M\left(\sum_{i=0}^{t-1}\frac{p_{H_1}(1-p_{H_1})^i}{1-(1-p_{H_1})^t}\, i\right)}{M}\right] = \sum_{i=0}^{t-1}\frac{p_{H_1}(1-p_{H_1})^i}{1-(1-p_{H_1})^t}\, i \end{aligned}$$

and

$$\begin{aligned} V(R_{H_1}) &= E\left(V\left(R_{H_1}|M>0\right)\right] + V\left[E\left(R_{H_1}|M>0\right)\right] = \\ & E\left[\frac{V\left(n_{(H_1 0)}|M>0\right)}{M^2}\right] + V\left[\frac{E\left(n_{(H_1 0)}|M>0\right)}{M}\right] = \\ & E\left[\frac{M\left(\sum_{i=0}^{t-1}\left(i - \sum_{i=0}^{t-1}\frac{p_{H_1}(1-p_{H_1})^i}{1-(1-p_{H_1})^t}\, i\right)^2 \frac{p_{H_1}(1-p_{H_1})^i}{1-(1-p_{H_1})^t}\right)}{M^2}\right] + 0 = \\ & E\left[\frac{1}{M}\right]\sum_{i=0}^{t-1}\left(i - \sum_{i=0}^{t-1}\frac{p_{H_1}(1-p_{H_1})^i}{1-(1-p_{H_1})^t}\, i\right)^2 \frac{p_{H_1}(1-p_{H_1})^i}{1-(1-p_{H_1})^t} \end{aligned}$$

Using the recursive formula related to inverse moments proposed in Zhao (2012) follows

$$V(R_{H_1}) = \frac{\sum_{j=1}^{N}\frac{P_0^{N-j}}{j} - P_0^n H_N}{1-P_0^N}\sum_{i=0}^{t-1}\left(i - \sum_{i=0}^{t-1}\frac{p_{H_1}(1-p_{H_1})^i}{1-(1-p_{H_1})^t}\, i\right)^2 \frac{p_{H_1}(1-p_{H_1})^i}{1-(1-p_{H_1})^t}$$

where H_N stands for the nth harmonic number, that is

$$H_N = \sum_{j=1}^{N} \frac{1}{j}$$

Notice that the expected value of R_{H_1} does not depend by the population size N while it depends from p_{H_1} and t: $E(R_{H_1})$ will decrease with p_{H_1} while it increases with t. Moreover, the simulations seem to show that for $p_{H_1} \to 0$ we have that $E(R_{H_1})$ approaches to the threshold $(t-1)/2$. On the other hand the variance of R_{H_1} depends on N. In fact for $N \to \infty$ we have that $V(R_{H_1})$ approaches zero and, as the expected value, it also increases with $1 - p_{H_1}$ and t.

	$p_{H_1} = 0.2$ $t = 5$	$p_{H_1} = 0.2$ $t = 10$	$p_{H_1} = 0.1$ $t = 5$	$p_{H_1} = 0.1$ $t = 10$
$N = 100$	$E(R_{H_1}) = 1.563$ $V(R_{H_1}) = 0.0281$	$E(R_{H_1}) = 2.797$ $V(R_{H_1}) = 0.0732$	$E(R_{H_1}) = 1.790$ $V(R_{H_1}) = 0.0489$	$E(R_{H_1}) = 3.647$ $V(R_{H_1}) = 0.1205$
$N = 1000$	$E(R_{H_1}) = 1.563$ $V(R_{H_1}) = 0.0028$	$E(R_{H_1}) = 2.797$ $V(R_{H_1}) = 0.0073$	$E(R_{H_1}) = 1.790$ $V(R_{H_1}) = 0.0048$	$E(R_{H_1}) = 3.647$ $V(R_{H_1}) = 0.0120$
$N = 10000$	$E(R_{H_1}) = 1.563$ $V(R_{H_1}) = 0.0003$	$E(R_{H_1}) = 2.797$ $V(R_{H_1}) = 0.0007$	$E(R_{H_1}) = 1.790$ $V(R_{H_1}) = 0.0005$	$E(R_{H_1}) = 3.647$ $V(R_{H_1}) = 0.0012$

Table 2.1: *Expected value and variance of R_{H_1} for different values of N, p_{H_1} and t.*

In Table 2.2 the expected value and variance for R_{H_1} corresponding on different values of p_{H_1} and t are reported. In Figure 2.1 are represented the realizations of the random variable R_{H_1} obtained in 1000 different data sets for each setting of p_{H_1} and N with $t = 3$. Analogously, figures 2.2, 2.3 and 2.4 represent the realizations of the same random variable with $t = 5, 7, 9$. The horizontal red line represents the failure threshold for R_{H_1}.

As we can see from the figures 2.1, 2.2, 2.3 and 2.4 the likelihood failure pathology persists also for high values of the population size ($N = 10000$) when the capture probability p_{H_1} is very low ($p_{H_1} = 0.05$) and when there are few capture occasions ($t = 3, 5$).

The likelihood failure problem is not overcome by using the unconditional likelihood. The unconditional MLE (UMLE) can be easily derived maximizing $L(N, \mathbf{p}_{\mathcal{H}_B})$ as a function of $\mathbf{p}_{\mathcal{H}_B}$ for N fixed so that once obtained $\hat{\mathbf{p}}_{\mathcal{H}_B}(N)$ one gets the profile likelihood $L_p(N) = L(N, \hat{\mathbf{p}}_{\mathcal{H}_B}(N))$ which can be in turn maximized as a function of N.

In Carle & Strub (1978) within the context of removal model it is pointed out the

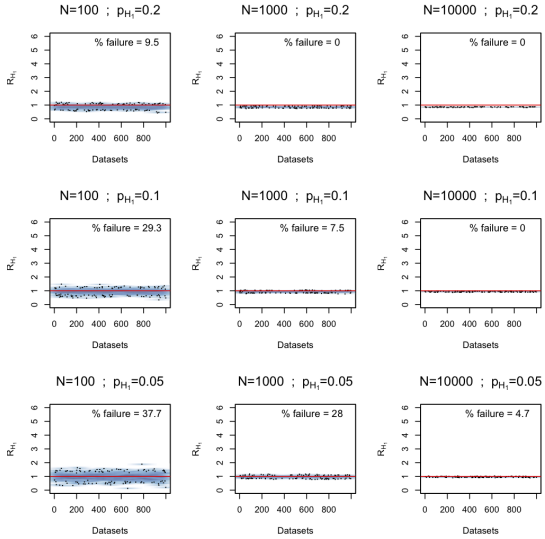

Figure 2.1: *Simulated values of* R_{H_1} *for different values of* N, p_{H_1} *and* $t = 3$

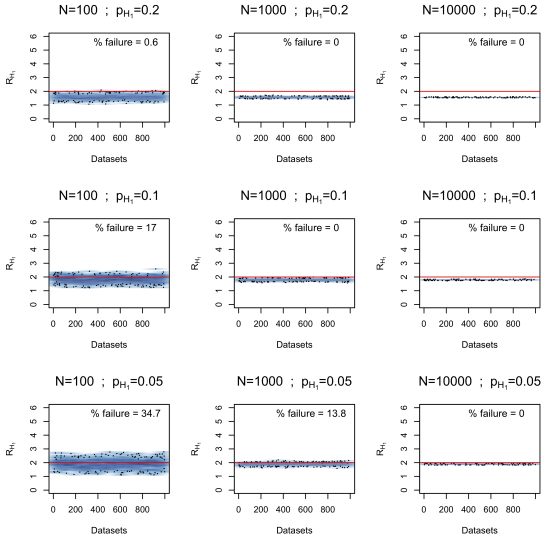

Figure 2.2: *Simulated values of* R_{H_1} *for different values of* N, p_{H_1} *and* $t = 5$

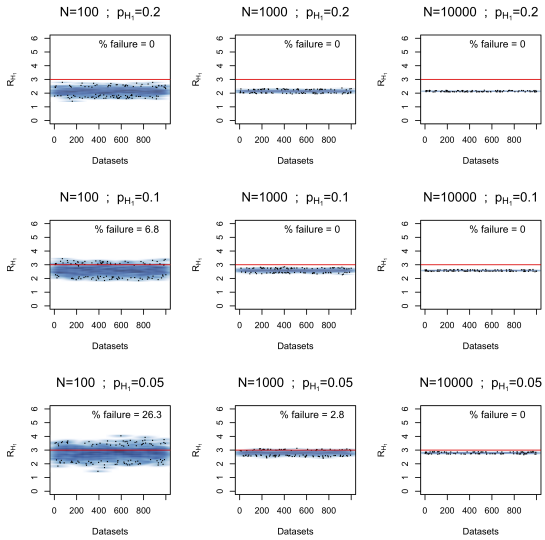

Figure 2.3: *Simulated values of* R_{H_1} *for different values of* N, p_{H_1} *and* $t = 7$

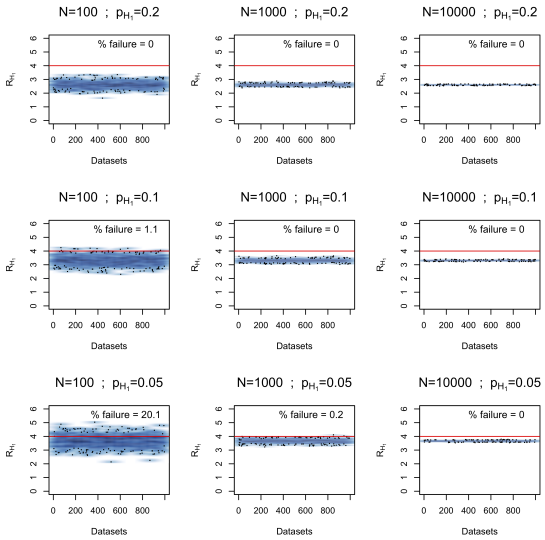

Figure 2.4: *Simulated values of* R_{H_1} *for different values of* N, p_{H_1} *and* $t = 9$

existence of the likelihood failure also for the unconditional likelihood approach providing the following conditions under which failure occurs

$$M(t-1) - n_{0p} \leq \frac{(M-1)(t-1)}{2} - 1 \Rightarrow \frac{n_{0p}}{M} \geq \frac{1}{2}(t-1) + \frac{t+1}{2M} \tag{2.7}$$

Notice that (2.7) is less restrictive than condition (2.6) in the removal model case and hence if failure occurs for the unconditional likelihood approach it occurs also for the conditional likelihood approach while vice-versa is not necessary true.
In order to completely overcome the likelihood failure problem for a removal study, Carle & Strub (1978) proposed to weight the likelihood function with a 2-parameter Beta distribution and then integrate out the nuisance parameter p. It is easy to understand that this procedure is equivalent to locating the posterior mode in a Bayesian approach with an improper non-informative uniform prior on N and a prior Beta distribution for p. In the original paper they show only by simulation how the integrated likelihood approach does not come across the problem of the likelihood failure.

2.3 Bayesian approach

Motivated by the solution proposed by Carle & Strub (1978) for the removal model we propose to extend the weighted likelihood approach as a fully Bayesian approach for the general class of behavioural models (2) and in particular for models in the class $\tilde{\mathcal{M}}$.
We will make use of Beta densities as convenient conjugate priors for each conditional probability $p_{H_b} \in \mathbf{p}_{\mathcal{H}_B}$. On the other hand we will consider a prior distribution on N as well. As reference recipes we have evaluated 4 non-informative prior distributions on N: Uniform, $1/N$ (Jeffreys'prior), $1/N^2$ and Rissanen's prior which represents a universal non-informative prior for discrete parameters (Rissanen 1983). Since we would like to pursue a fully Bayesian approach we need to verify whether the first two improper priors lead to proper posterior distributions. If this is the case we can fully exploit alternative summaries of the posterior distribution on N. In particular we will consider as alternative summaries the mean, the median, the mode and a minimizer of a specific loss function $\mathcal{L}$ connected with the Relative Mean Square Error (RMSE) as in Tardella (2002)

$$\mathcal{L}(a, N) = \left(\frac{a}{N} - 1\right)^2$$

Let $\pi(N, \mathbf{p}_{\mathcal{H}_B})$ be the joint prior distribution on the whole parameter vector $(N, \mathbf{p}_{\mathcal{H}_B})$ such that

$$\pi(N, \mathbf{p}_{\mathcal{H}_B}) = \pi(N) \times \prod_{b=1}^{B} \pi(p_{H_b}) \propto \pi(N) \times \prod_{b=1}^{B} p_{H_b}^{\alpha_b - 1}(1 - p_{H_b})^{\beta_b - 1} \quad (2.8)$$

Hence, given (2.3) and (2.8) the joint posterior distribution for model $M_{\mathcal{H}_B}$ is

$$\pi(N, \mathbf{p}_{\mathcal{H}_B}|\mathbf{X}) \propto L(N, \mathbf{p}_{\mathcal{H}_B})\pi(N, \mathbf{p}_{\mathcal{H}_B}) \propto$$
$$\pi(N)\binom{N}{M} p_{H_1}^{n_{(H_1 1)}+\alpha_1-1}(1-p_{H_1})^{n_{(H_1 0)}+t(N-M)+\beta_1-1}\prod_{b=2}^{B} p_{H_b}^{n_{(H_b 1)}+\alpha_b-1}(1-p_{H_b})^{n_{(H_b 0)}+\beta_b-1}$$

The choice of Beta densities as prior distributions for the conditional probability parameters makes the marginal posterior distribution of N available in closed form up to a normalizing constant as follows

$$\begin{aligned}\pi(N|\mathbf{X}) &= \int_0^1 \dots \int_0^1 \pi(N, \mathbf{p}_{\mathcal{H}_B}|\mathbf{X}) d\mathbf{p}_{\mathcal{H}_B} \\ &\propto \pi(N)\frac{N!}{(N-M)!} B(n_{(H_1 1)}+\alpha_1, t(N-M)+n_{(H_1 0)}+\beta_1) \quad (2.9)\end{aligned}$$

where $B(,)$ is the Beta function. The posterior marginal distribution of N in closed form as in (6) makes it easy to compute quickly all the posterior summaries. We will show how the choice of the prior distribution on N has a relevant impact on the posterior summaries while the sensitivity with respect to the choice of the parameters of the Beta distribution is less relevant. In the following we consider a uniform density on p_{H_b} corresponding to Beta parameters $\alpha_b = \beta_b = 1$, for $b = 1, \dots, B$. As preliminary step we formally verify whether the choice of improper prior distributions on N such as $\pi(N) \propto 1$ and $\pi(N) \propto 1/N$ leads to a proper marginal distribution on N.

Lemma

Consider a generic model within the class $\tilde{\mathcal{M}}$ parametrized in terms of $p_{\mathcal{H}_B}$. If one chooses independent uniform priors for all its components and a noninformative prior on N with probabilities $\pi(N) \propto 1/N^r$ the Bayes rule always yields a proper posterior distribution for any $r > 0$ while for $r = 0$ the condition $n_{H_1 1} > M$ suffices.

Proof – Considering $\pi(N) \propto 1/N^r$ the posterior marginal distribution of N is proportional to

$$\pi(N|\mathbf{X}) \propto \frac{1}{N^r}\frac{\Gamma(N+1)}{\Gamma(N-M+1)}\frac{\Gamma(t(N-M)+n_{(H_1 0)}+1)}{\Gamma(t(N-M)+n_{(H_1 0)}+n_{(H_1 1)}+2)}$$

Using the inequalities

$$(2\pi)^{\frac{1}{2}} x^{x-\frac{1}{2}} \exp(-x) \leq \Gamma(x) \leq (2\pi)^{\frac{1}{2}} x^{x-\frac{1}{2}} \exp(-x+1/12x)$$

one gets

$$\frac{\Gamma(N+1)}{\Gamma(N-M+1)} < \mathcal{O}(N^M) \; ; \; \frac{\Gamma(t(N-M)+n_{(H_1 0)}+1)}{\Gamma(t(N-M)+n_{(H_1 0)}+n_{(H_1 1)}+2)} < \mathcal{O}(N^{-(n_{(H_1 1)}+1)})$$

Hence, $\pi(N|\mathbf{X}) < \mathcal{O}(N^{M-(r+n_{(H_1 1)}+1)})$ which corresponds to a proper pmf if and only if $M-(n_{(H_1 1)}+r+1) < -1$. We have to distinguish 2 cases: when $n_{(H_1 1)} > M$ the marginal posterior distribution on N is always proper for any $r \geq 0$ while when $n_{(H_1 1)} = M$ this is true only for $r > 0$ $\diamondsuit$

From the proof of the previous lemma one can easily argue formally that the Bayes rule always provides an eventually vanishing function in the numerator of (2.9) for $N \to \infty$. This important result shows that if one makes inference on N maximizing

$$W(N|\mathbf{X}) = \pi(N)\frac{N!}{(N-M)!}\, B(n_{(H_1 1)}+1, t(N-M)+n_{(H_1 0)}+1) \tag{2.10}$$

one will never get unbounded estimates for the finite population size no matter what improper prior is chosen within the class of $1/N^r$ for $r \geq 0$. Hence the Bayesian approach can never provide unbounded estimates in the following sense.

Corollary

Consider a generic model within the class $\tilde{\mathcal{M}}$ parametrized in terms of $\mathbf{p}_{\mathcal{H}_B}$. If one chooses independent uniform priors for all p_{H_b} components and a noninformative prior on N with probabilities $\pi(N) = 1/N^r$ then there exists $\hat{N}_{mode} < \infty$ such that $W(\hat{N}_{mode}|\mathbf{X}) \geq W(N|\mathbf{X})$ for any N.

Notice that $\hat{N}_{mode}$ is the mode of the posterior distribution only in those cases where it is well defined otherwise it can be considered only as a weighted likelihood. Hence we claim that from a theoretical inferential point of view the Bayesian approach should be regarded in this context as a favorite inferential tool since it always yields valid inference. Unfortunately it is not easy to get explicit formulas to determine how likely the occurrence of the likelihood failure is. Of course that will depend on the true model and parameter configurations. In the following section we investigate the issue with a little simulation study with replicated data from the same model. Moreover we will show that even when we remove from the analysis those data which yield likelihood failure the comparative performance of Bayesian output versus conditional maximum likelihood is still always in favor of the former.

2.4 Simulation study

In order to evaluate the comparative performance of the Bayesian approach with respect to the classical approach based on conditional likelihood we propose a small simulation study. We consider the set of simulation trials described in Table 2.4. The true population size is $N = 100$ and the number of trapping occasions is $t =$

5. We evaluate three different kinds of behavioural models within the extended Markovian structure $M_{c_k b}$ following the ideas in Yang & Chao (2005) to account for both enduring and ephemeral effects. Indeed, Markov order is restricted to 2 and we have also excluded $M_{c_1 b}$ and $M_{c_2 b}$ from consideration since they yield inference on N which is identical to model M_b for the reasons we have explained in the previous section. The true (conditional) capture probability parameters for the different simulation trials are chosen so that they correspond to different degrees, from medium-high to medium-low, of expected capture sample coverage defined as the fraction of distinct individuals observed during the t trapping stages, in symbols

$$\frac{E[M]}{N} = 1 - P_0$$

Trial	**Model**	**Probability parameters**	$E[M]/N$
$Tr.1$	M_b	$p = 0.2; r = 0.4$	0.67
$Tr.2$	M_b	$p = 0.1; r = 0.3$	0.41
$Tr.3$	M_{c_1}	$p_{(0)} = 0.2; p_{(1)} = 0.4$	0.67
$Tr.4$	M_{c_1}	$p_{(0)} = 0.1; p_{(1)} = 0.3$	0.41
$Tr.5$	M_{c_2}	$p_{(00)} = 0.2; p_{(10)} = 0.3; p_{(01)} = 0.35; p_{(11)} = 0.4$	0.67
$Tr.6$	M_{c_2}	$p_{(00)} = 0.1; p_{(10)} = 0.2; p_{(01)} = 0.3; p_{(11)} = 0.4$	0.41

Table 2.2: *Parameter configurations for simulation experiments. For each parameter configuration* $K = 1000$ *datasets have been simulated.*

Notice that we have used for each simulated trial the same sequence of pseudo-random numbers so that the observed number of distinct units in each trial is the same when the probabilities p, $p_{(0)}$, and $p_{(00)}$ are the same. To summarize the posterior distribution of the main parameter of interest N we consider the usual mean, median and mode together with the posterior loss minimizer for the loss function described in Section 4

$$m_R = \arg\min_a E_{\pi(N|\mathbf{x})}(\mathcal{L}(a, N)).$$

In Table 2.4 we report the root of the relative mean square error (RMSE) of the estimates of N based on simulations from the correct model. RMSE is evaluated empirically on the basis of $K = 1000$ replicated datasets for each trial. As we can see the Bayesian approach outperforms the CMLE and UMLE in terms of RMSE. Indeed the occurrence of likelihood failure is reported in the last lines of Table 2.4 as a percentage of the K datasets. In reporting the estimated RMSE the * sign

denotes the presence of likelihood failure so that the RMSE is indeed computed as restricted RMSE conditioning on the absence of failure. This means that RMSE is computed conditioning only on datasets which lead to a finite value of $\hat{N}_{CMLE}$ and $\hat{N}_{UMLE}$. Table 2 allows to assess the comparative performance of alternative choices as far as $\pi(N)$ is concerned. We remark that the choice has some impact on the frequentist performance. Similarly, the choice of posterior summary has a remarkable effect on the precision of the resulting estimator. Our simulations show that the combination of $\pi(N)$ and summary which produces a better performance corresponds to either one of posterior mode and m_R combined with $1/N^2$. Overall the option m_R with Rissanen shows a more robust behaviour even when they are not the best combination since its RMSE is always close to the best one.
Notice also that the $\hat{N}_{CMLE}$ seems to be more accurate than $\hat{N}_{UMLE}$ in trial 1,2,6 but this is due to the fact that the RMSE are restricted RMSE computed considering different subsets of the K datasets.

We have also considered the performance of alternative approaches with respect to interval estimators. In Table 3 we report the actual percentage of trials in which the 95% interval estimates covered the true value of N and also the average length of the intervals. For the classical approach $1-\alpha$ confidence intervals for the population size are obtained through the profile log-likelihood as (N^-, N^+) where N^- and N^+ are the two roots of the following equation

$$2(\log(L_p(\hat{N})) - \log(L_p(N))) = z^2_{\alpha/2}$$

where $z_{\alpha/2}$ is the $\alpha/2$ quantile of the standard normal and L_p is the profile likelihood.

As in Table 2 the * sign denotes the presence of likelihood failure while the $ sign warns that the actual average length is greater than the reported value since we have arrested the root finding to an upper-bound $N^+_{upper} = 10000$. In those cases we have set $N^+ = N^+_{upper}$. In fact in some dataset, although the failure condition is not met the flatness of the profile likelihood prevent us from locating the root N^+ before N^+_{upper}. For the Bayesian approach we have computed the HPD credible set with the same nominal $1-\alpha$ posterior probability value. The prior $\pi(N) = 1/N^2$ leads to the smallest interval estimates, but the actual coverage is not always sufficiently close to the level $1-\alpha$ desired for a frequentist match. For trial 2, 4 and 6, characterized by a moderately low sample coverage $E[M]/N$, the coverage of the Bayesian interval estimator corresponding to $\pi(N) = N^{-2}$ is significantly lower than 95% while this is not true for the Rissanen prior. Even for interval estimate purposes Rissanen's prior represents a good compromise: the average length is reasonably small and the coverage is appropriately close to the nominal frequentist match.

Prior	**Estimator**	Tr. 1	Tr. 2	Tr. 3	Tr. 4	Tr. 5	Tr. 6
$1/N$	Mean	0.999	1.400	0.188	1.265	0.535	2.787
	Median	0.378	0.435	0.163	0.589	0.286	0.594
	Mode	0.173	0.391	0.137	0.288	0.163	0.330
	m_R	0.220	**0.306**	0.145	0.289	0.194	0.288
$1/N^2$	Mean	0.374	0.356	0.163	0.463	0.271	0.454
	Median	0.216	0.323	0.146	0.313	0.195	0.295
	Mode	**0.167**	0.421	**0.132**	0.286	**0.150**	0.345
	m_R	**0.167**	0.350	0.134	**0.262**	0.159	0.291
Rissanen	Mean	0.688	0.807	0.177	0.895	0.410	1.109
	Median	0.293	0.342	0.155	0.445	0.241	0.407
	Mode	0.170	0.407	0.135	0.285	0.156	0.330
	m_R	0.194	0.327	0.140	0.273	0.178	**0.280**
	CMLE	1.149*	1.284*	0.176	0.642*	0.337*	0.652*
	% of $\hat{N}_{CMLE} < \infty$	(99.3%)	(80.5%)	(100.0%)	(98.2%)	(99.8%)	(83.8%)
	% of $\hat{N}_{CMLE} = \infty$	(0.7%)	(19.5%)	(0.0%)	(1.8%)	(0.2%)	(16.2%)
	UMLE	1.341*	1.835*	0.166	0.592*	0.280*	0.757*
	% of $\hat{N}_{UMLE} < \infty$	(99.7%)	(86.3%)	(100.0%)	(98.2%)	(99.8%)	(85.2%)
	% of $\hat{N}_{UMLE} = \infty$	(0.3%)	(13.7%)	(0.0%)	(1.8%)	(0.2%)	(14.8%)

Table 2.3: *Simulated data: estimated $\sqrt{RMSE}$ based on 1000 replicated datasets for each trial. For each simulation setting (column) bold values highlight the best performing estimation method and the corresponding $\sqrt{RMSE}$.*

Now we deal with the model selection issue. Consider the simulation setting proposed in Tab. 2.4. The candidate models are M_b, M_t, M_{c_1}, M_{c_2}, M_{c_1b}, and M_{c_2b}. AIC and marginal likelihood are used as selection criteria for classical and Bayesian approach respectively. Three different priors have been implemented and the corresponding results compared: Rissanen, $1/N$ and uniform. As summary statistics we will consider the posterior mode and the expected loss minimizer m_R. The bar graphs in Figure 2.5 represent the number of times that a model is chosen. Moreover, the empirical coverage and $RMSE$ are reported in Table 2.5. Notice that, as discussed above, the uniform prior does not ensure the propriety of the posterior distribution. It is equivalent to an integrated likelihood approach with uniform weight function on the probabilities involved in the model. Hence we will consider the mode only as summary statistic. As expected, in trials 1,2,3 and 4 both selection criteria get the correct model most of the times although ML selects the true model more

	Interval Estimate	**Coverage**	**Average length**
Tr.1	Bayes $(1/N^2)$	95.3%	120.87
	Bayes (Rissanen)	96.0%	194.46
	Classical PLI	95.1%*$	$\geq$ 2388.94*$
Tr.2	Bayes $(1/N^2)$	88.0%	163.94
	Bayes (Rissanen)	92.8%	342.04
	Classical PLI	94.5%*$	$\geq$ 7201.63*$
Tr.3	Bayes $(1/N^2)$	94.7%	57.07
	Bayes (Rissanen)	95.1%	59.94
	Classical PLI	94.5%	69.43
Tr.4	Bayes $(1/N^2)$	90.6%	144.4
	Bayes (Rissanen)	97.3%	204.13
	Classical PLI	94.5%*	488.80*
Tr.5	Bayes $(1/N^2)$	94.8%	81.92
	Bayes (Rissanen)	95.7%	97.31
	Classical PLI	94.9%*	175.10*
Tr.6	Bayes $(1/N^2)$	89.8%	172.67
	Bayes (Rissanen)	93.0%	319.40
	Classical PLI	97.2%*$	$\geq$ 855.24*$

Table 2.4: *Simulated data: empirical coverage and average length in simulated data of alternative interval estimates with nominal confidence level 0.95 and posterior probability 0.95 respectively*

often than AIC. This is no longer true in settings 5 and 6 where the parameter configurations of the transition probabilities for model M_{c_2} are not very different from a M_{c_1} setting and with only five capture occasions both AIC and marginal likelihood are not able to select correctly the true model preferring the more parsimonious M_{c_1}. As shown in the results in Table 2.5 Rissanen's prior and $1/N$ lead to very similar $RMSE$ yielding the best performances. However, using Rissanen's prior interval estimates do not always achieve the desired nominal coverage of 95% especially for simulation settings where the expected value of the number of observed distinct units is low (Trial 2,4,6). Notice also that the weighted likelihood approach (uniform prior) yields interval estimates with almost perfect nominal coverage of 95% although it does not provide as accurate point estimates as the fully Bayesian analyses.

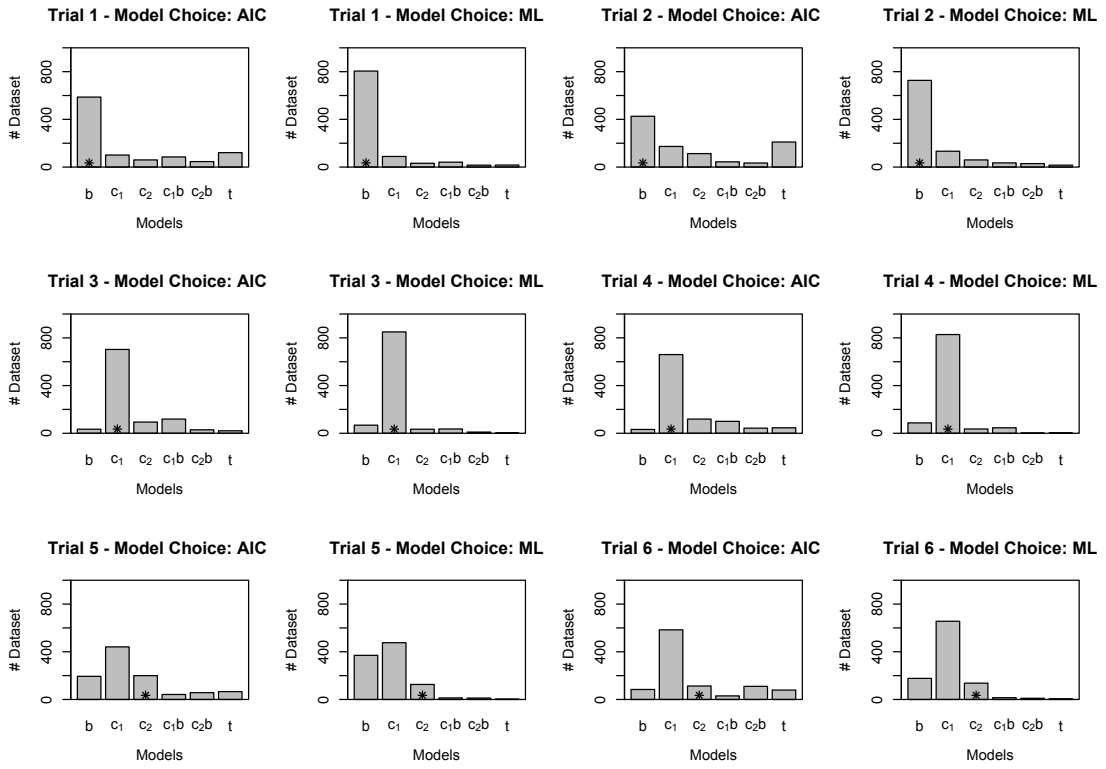

Figure 2.5: *Model selection in simulated data: frequency histograms showing the number of times that a specific model is selected as the best one in terms of AIC and marginal likelihood criteria respectively. The * sign denote the true model used to simulate data.*

	UMLE		**Uniform**		$1/N$			**Rissanen**		
Trial	$RMSE$	%	$RMSE_{mode}$	%	$RMSE_{mode}$	$RMSE_{mR}$	%	$RMSE_{mode}$	$RMSE_{mR}$	%
1	1.16	79.8	0.21	94.4	0.18	0.22	91.8	0.17	0.20	90.5
2	1.52	65.1	0.37	94.9	0.39	0.32	88.9	0.41	0.34	84.3
3	1.10	90.0	0.21	96.2	0.16	0.19	95.3	0.15	0.17	94.8
4	0.88	90.3	0.44	97.2	0.31	0.30	93.2	0.31	0.29	90.3
5	1.15	86.3	0.23	95.0	0.17	0.22	92.9	0.16	0.22	91.8
6	1.31	82.6	0.36	94.3	0.33	0.30	87.3	0.34	0.30	82.6

Table 2.5: *Model selection in simulated data: root mean square error of point estimates and empirical coverage of interval estimates with nominal confidence level 0.95 and posterior probability 0.95 respectively. Notice that point and interval estimates are derived after model selection.*

2.5 Real data

We reanalyze the Great-Copper butterfly dataset originally studied in Ramsey & Severns (2010) to support the use of more flexible behavioural models to account for possibly decreasing/increasing recapture probability patterns likely to occur after the first capture of each unit. It is supposed that butterflies are subject to a change of behaviour which persists with different intensity until the end of trapping stages. In Ramsey & Severns (2010) three alternative models denoted with M_p, M_{pt}, M_{pb} are proposed and referred to as *persistence models* (see the original paper for a more detailed description of these models). Indeed they do not belong to the class $\mathcal{M}$ of conditional probability models within the framework proposed in Farcomeni (2011). This persistence phenomenon can be considered as a *trap-happiness* response and it can be justified from the fact that butterflies are used to return to the same place where the food is in great quantity. Analogously, researchers are used to return to the same place where they find butterflies. The same dataset is also reviewed in Farcomeni (2011) to show that the class $\mathcal{M}$ is flexible enough to accommodate behavioural models which fit the same data better. The experiment is made of $t = 8$ trapping occasions and the number of distinct butterflies captured during all trapping stages is $M = 45$. In Table 3 we report only the observed complete capture histories associated with the respective frequencies.

We fit several models based on different partitions of the set H some of which correspond to alternative versions of $M_{c_k b}$. Model M_L originally proposed in Farcomeni (2011) considers a 3-rd order Markov-chain-like structure where capture probabilities depend only on the previous three occasions but, differently from the full model M_{c_3} which contains $2^3 = 8$ probability parameters, it considers only 2 parameters corresponding to the following (bi)partition $\mathcal{H}_2(M_L) = \{H_1, H_2\}$ such that

$$\begin{cases} H_1 = \{(), (0), (10), (x_1, \ldots, x_{j-4}, 0, 0, 0), \\ \qquad (x_1, \ldots, x_{j-4}, 1, 0, 0), (x_1, \ldots, x_{j-4}, 0, 1, 0), \\ \qquad (x_1, \ldots, x_{j-4}, 1, 1, 0), (x_1, \ldots, x_{j-4}, 0, 0, 1)\} \\ \qquad\qquad\qquad\qquad \forall (x_1, \ldots, x_{j-4}) \in \mathcal{X}^{j-4} \quad ; \quad \forall j \geq 4 \\ H_2 = H \setminus H_1 \end{cases}$$

The parameter p_{H_1} corresponding to the first partition identifies a vanishing behavioural effect which occurs if the unit is not captured in the most recent occasion, or captured only once in the last three occasions.

In Table 4 we display point and interval estimates at level 95% of population size N derived with both classical and Bayesian approach. As described in Section 3 the confidence intervals are built considering the normal approximation of the profile

History	Butterflies
0 0 0 0 0 0 0 1	3
0 0 0 0 0 0 1 0	3
0 0 0 0 0 0 1 1	1
0 0 0 0 0 1 0 0	4
0 0 0 0 1 0 0 0	4
0 0 0 0 1 0 0 1	1
0 0 0 0 1 1 0 0	1
0 0 0 1 0 0 0 0	3
0 0 0 1 0 1 0 0	2
0 0 0 1 1 0 0 0	1
0 0 1 0 0 0 0 0	4
0 1 0 0 0 0 0 0	5
0 1 0 0 0 0 1 0	1
0 1 0 1 0 1 1 0	1
0 1 1 0 0 0 0 0	1
0 1 1 0 1 0 0 0	1
0 1 1 1 1 0 1 1	1
1 0 0 0 0 0 0 0	5
1 1 0 0 0 0 0 0	1
1 1 1 0 0 0 0 0	1
1 1 1 1 1 1 0 0	1

Table 2.6: *Great Copper Butterfly data: frequencies of observed capture histories*

log-likelihood while for the Bayesian approach we have proposed the HPD interval. Furthermore in Table 4 in order to drive model selection we report both the AIC index and the log-marginal likelihood associated to each model.
In order to get insights on the pattern of behavioural effects we look at the posterior distribution of $p_{H_2} - p_{H_1}$ for models which involve $\mathbf{p}_{\mathcal{H}_2} = (p_{H_1}, p_{H_2})$ as nuisance parameter.

In Figure 2.6 we display the posterior densities of $p_{H_2} - p_{H_1}$ for models M_b, M_{c_1} and M_L. Model M_b which considers only the classical enduring effect to capture provides evidence of trap-shyness. In fact the distribution $p_{H_2} - p_{H_1} = r - p$ is well concentrated almost entirely below the value zero. On the other hand both models M_{c_1} and M_L present trap-happiness effect ($p_{H_2} - p_{H_1} > 0$) more consistent with the underlying biological assumptions.
Following the recommendation suggested by our simulation study we have used Ris-

Model	**# parameters**	**Approach**	$\hat{N}$	(N^-, N^+)	**AIC**	**log-ML**
M_0	1+1	CMLE	65	(52,86)	336.80	
		Bayesian	63	(51,82)		-174.68
M_t	1+8	CMLE	64	(52,85)	350.84	
		Bayesian	59	(49,72)		-187.14
M_b	1+2	CMLE	67	(48,223)	342.77	
		Bayesian	63	(46,135)		-176.57
M_{c_1}	1+2	CMLE	96	(64,181)	328.92	
		Bayesian	88	(58,151)		-169.63
M_{c_2}	1+4	CMLE	176	(78,896)	326.26	
		Bayesian	117	(59,374)		-169.51
M_{c_3}	1+8	CMLE	174	(69,2315)	330.16	
		Bayesian	106	(53,419)		-175.83
$M_{c_1 b}$	1+3	CMLE	67	(48,223)	329.24	
		Bayesian	63	(46,135)		-170.91
$M_{c_2 b}$	1+5	CMLE	67	(48,223)	324.50	
		Bayesian	63	(46,135)		-169.93
$M_{c_3 b}$	1+9	CMLE	67	(48,223)	328.19	
		Bayesian	63	(46,135)		-173.62
M_L	1+2	CMLE	90	(63,152)	324.01	
		Bayesian	84	(58,133)		-166.91
M_p	1+2	CMLE	97	(70,215)	328.92	
M_{pt}	1+9	CMLE	64	(54,103)	339.46	
M_{pb}	1+2	CMLE	69	(60,1006)	330.16	

Table 2.7: *Great Copper Butterfly data: AIC, log marginal likelihood, point and interval estimates*

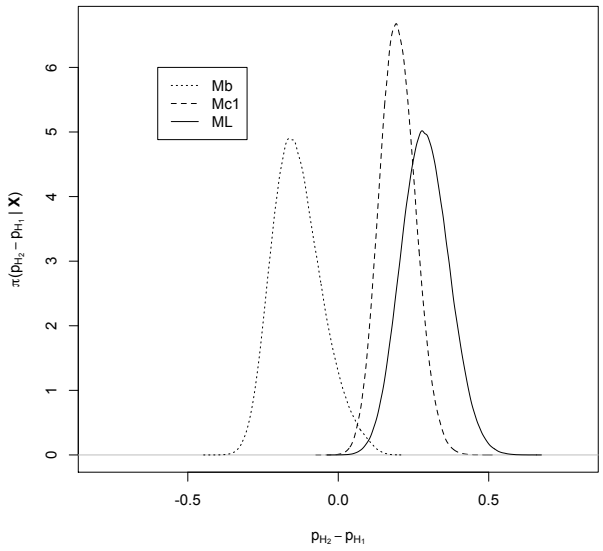

Figure 2.6: *Great Copper Butterfly data: posterior distribution of* $p_{H_2} - p_{H_1}$ *for models* M_b, M_{c_1} *and* M_L.

sanen's prior as prior distribution of N since it yields more convincing results than those provided by $\pi(N) = 1/N^2$. From Table 2.7 one can observe how the Bayesian approach always yields estimates of the population size N which are smaller than CMLE. This is indeed expected from the fact that the Bayesian approach makes full use of the (integrated) unconditional likelihood and the well known monotonicity properties with respect to the estimation based on the conditional likelihood (Sanathanan 1972). From the kind of forest plot in Figure 2.5 it is also easy to appreciate that Bayesian approach provides narrower and more stable interval estimates than those provided by a frequentist approach based on the profile likelihood corresponding to comparable $1 - \alpha$ levels. In particular model M_{c_2} and M_{c_3} yield very wide classical confidence intervals which reflect the relative flatness of the profile likelihood. In Table 2.7 we report for completeness the results of Ramsey & Severns (2010) for their proposed models M_p, M_{pt} e M_{pb} to highlight how instability of classical estimators based on CMLE together with wide confidence intervals may be present also in behavioural models which are outside the $\tilde{\mathcal{M}}$ class of models unraveling that likelihood flatness problems lurks behind.
Notice also that the AIC index and the log marginal likelihood (log-ML) agree on the choice of L_2 as the best model. However, the log-ML gives stronger support than AIC to more parsimonious Markovian models such as M_{c1} and M_{c1b} while it rather penalizes M_{c3} and M_{c3b} which include a higher number of parameters.

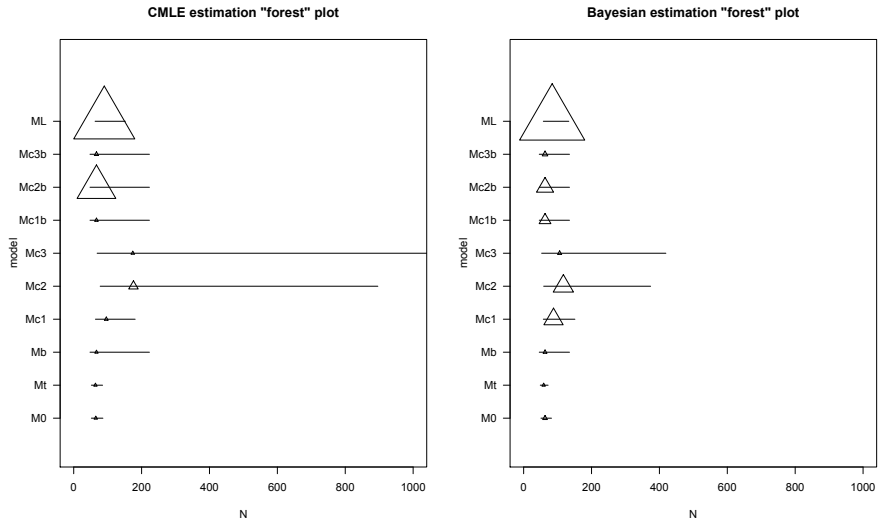

Figure 2.7: *Great Copper Butterfly data: forest plots for interval estimates of N. Triangles locate point estimates while their sizes are proportional to the amount of comparative evidence evaluated either by AIC or by log-ML*

2.6 Final remarks

In order to understand behavioural patterns in capture recapture experiments we have focussed on a general class of models following the approach of Farcomeni (2011). Instead of adopting more conventional tools for categorical (binary) data it relies on the reparameterization of the joint probability of the multivariate binary outcome corresponding to the entire individual capture history in terms of subsequent conditional probabilities. This is in the same spirit of the so-called transitional model reviewed in Zeng & Cook (2007). Our choice is appropriate since we believe that a behavioural pattern is more easily understood and formalized in terms of conditional probabilities.

We have then pointed out that with the conditional likelihood approach a possible unbounded estimate of the parameter of interest can occur and such pathological inferential feature is indeed shared by a large class of behavioural models both with enduring and ephemeral effects. This phenomenon is rather neglected in the literature since most of the analyses are based on conditional likelihood (Huggins & Hwang 2011).

In the literature there are other classes of capture-recapture models where likelihood

failure may occur. In particular some parametric and nonparametric heterogeneity models labelled as M_h have been considered in Mao & You (2009) following some critical remarks raised by Link (2003) on model identifiability. They showed with simulated examples similar likelihood pathologies (see Table 5 therein and related comments). However, as said in the introduction, we opted for distinguishing the pathologies derived by the heterogeneity from those due to behavioural effect modeling.

Hence focussing on classes of behavioural models with no heterogeneity such as those derived from the approach of Farcomeni (2011) we have characterized with the subclass $\tilde{\mathcal{M}}$ some models and conditions under which likelihood failure occurs and we have shown that even when there is no likelihood failure the inferential output can be very large and unstable. On the other hand in a very flexible model framework for behavioural patterns we have shown that a fully Bayesian approach is a viable solution which brings a two-fold beneficial effect on inference: i) a simple conjugate structure with closed form expressions for the marginal posterior probabilities $\pi(N|X)$ up to a normalizing constant ii) the complete overcome of unbounded inference under any observed dataset. Since Bayesian inference requires the specification of prior distributions on the unknown parameters we have investigated the sensitivity of the analysis with respect to few alternative default priors using their frequentist properties as performance criterion. Our analysis strongly supports the use of a fully Bayesian approach within the class of models $\tilde{\mathcal{M}}$ based on grouping of the conditional probabilities in equivalence classes. As default choice we advocate the use of uniform priors on the conditional probability parameters and a Rissanen prior on the integer parameter representing the unknown population size N. In our simulations this choice provided improved inference in terms of reduced relative mean square error and shorter interval estimates in the presence of equivalent frequentist coverage. This remains true, although at a lesser extent, when the comparison with unconditional MLE is considered.

An anonymous referee suggested the possibility of using a generalized log-linear parameterization as in Lang (1996) to get Farcomeni's model framework as a particular instance. However we found the implementation of such idea not straightforward and we will look forward to further investigation on that. Indeed we point out the possibility of using a logistic reparameterization of the probability of each binary outcome of the capture history to derive unconditional MLE. In fact one can consider the logit of the conditional probability of each binary outcome regressed as a suitable function of the previous partial capture history. When such function corresponds to a categorical covariate assuming levels corresponding to each equivalence class the derivation of the unconditional MLE can be easily carried out by maximizing the profile likelihood of N. In fact for each value of N one can augment the

observed capture histories with $N - M$ histories corresponding to units which were not observed and obtain the profile likelihood corresponding to N from the standard output of GLM routines of any statistical software. A similar logistic model structure has been previously sketched in Huggins (1989) and Alho (1990) although the focus there was in developing conditional likelihood estimators in the presence of individual covariates different from partial capture histories.

We believe that the generality of the pathological features of classical likelihood analysis (CMLE and UMLE) of behavioural capture recapture models suggests a wider use of Bayesian alternative analysis even in those more realistic and complex frameworks such as, for instance, those developed in Bartolucci & Pennoni (2007) where latent Markov structure is embedded to model more flexibly ephemeral effects and heterogeneity of individual capture probability. Another critical point that is not addressed here is related to possible relaxation of the independence hypotheses among units (Fattorini et al. 2007).

Chapter 3

Behavioural modeling via scaling of partial capture history

As shown in the previous chapter the framework proposed in Farcomeni (2011) is a very general tool to model behavioural effects to capture embedding also, as special cases, several classical models already proposed in the literature such as M_b, M_t (Otis et al. 1978), M_{c_k}, $M_{c_k b}$ (Yang & Chao 2005 , Farcomeni 2011).
In this section we propose a new flexible expedient to model and interpret the behavioural effect to capture which cannot be recovered as special case by already available model frameworks. We take once again the approach of conditioning to the previous partial capture histories as a natural way of keeping track of the sequential behavioural changes. Our idea in order to allow a meaningful behavioural effect to capture in the model relies on a suitable ordering and scaling of the sequences of progressive individual partial capture histories to be summarized in a numerical time-dependent individual covariate evolving longitudinally through the whole capture sequence. The proposed ordering is based on the binary representation of integers used in Lloyd & Frommer (2008) to sort out longitudinal binary outcomes in the context of a multiple-screening test. We show how an appropriate rescaling of such representation can be fruitfully used as a suitable quantitative covariate to be embedded in a logistic regression or any other generalized linear model (GLM) framework. Besides, we will show how this strategy can be related to the idea of fitting variable order Markov structures (Buhlman et al. 2007) possibly recovering again, as special cases, some classical behavioural models such as M_b, M_{c_k} and $M_{c_k b}$. Obviously this idea can be extended in all longitudinal studies which are concerned with binary longitudinal predictor variables and are not confined in a capture-recapture context only. Indeed, many longitudinal studies use binary covariates which are observed across time such as the weekly presence (1) or absence

(0) of a specific disease in an epidemiological survey.

3.1 Meaningful numeric covariate representation of longitudinal binary outcomes

The starting point is to consider, similarly to Huggins (1989) and Alho (1990), a logistic regression model viewing each capture occurrence of unit i at occasion j as a binary outcome whose probability can be modelled as a function of a synthetic explanatory variable $z_{ij} = q(x_{i1}, \ldots, x_{ij-1})$ associated to the previous (partial) capture history. Formally this can be expressed as follows

$$\begin{aligned} \text{logit}\left(p_j(x_1, \ldots x_{j-1})\right) &= \log\left(\frac{Pr(X_{ij}=1|x_{i1}=x_1, \ldots x_{ij-1}=x_{j-1})}{1-Pr(X_{ij}=1|x_{i1}=x_1, \ldots x_{ij-1}=x_{j-1})}\right) \\ &= r(q(x_1, \ldots x_{j-1})) \\ &= r(z_{ij}) \end{aligned} \tag{3.1}$$

Our first idea is to consider a simple linear logistic regression for the probability of each capture event X_{ij},

$$\text{logit}\left(p_j(x_1, \ldots x_{j-1})\right) = \alpha + \beta z_{ij} \qquad \forall i \, \forall j \tag{3.2}$$

where z_{ij} is a suitable numeric summary or quantification of partial capture history $(x_{i1} \ldots, x_{ij-1})$.

We link the partial capture history denoted with $\mathbf{x} = (x_1, \ldots, x_r)$ to the notation used in the previous chapter where a capture history is represented as a binary string taking values in $H = \cup_{r=0}^{t-1} \mathcal{X}^r$. The length of a partial capture history $\mathbf{x} \in H$ is denoted with $l_{\mathbf{x}}$. Any partial capture history can be transformed into an integer number z using the string as its binary representation. To simplify the notation the unit index i will be omitted in the following when it is not needed. According to the natural and intuitive interpretation of grading a behavioural effect so that the occurrence of trapping in the last occasions has a greater impact on the future capture probability than those occurred in the previous ones we proceed to appropriately reverse the usual representation and consider the following binary representation

$$f(\mathbf{x}) = f(x_1, \ldots, x_{l_{\mathbf{x}}}) = \sum_{j=1}^{l_{\mathbf{x}}} x_j 2^{j-1} \in \{0, 1, 2, ..., 2^{l_x} - 1\}$$

where we assume that the partial capture history has length $l_{\mathbf{x}} \geq 1$. Conventionally, we set $f(\mathbf{x}) = 0$ for the empty binary sequence of length zero corresponding to

$\mathbf{x} = ()$.
However, note that according to the length $l_{\mathbf{x}}$ of the binary string one gets a different range of integers. Hence, in order to obtain a potentially continuous covariate in a fixed range to be used as a synthetic representation of the past history we rescale the range $[0, 2^{l_{\mathbf{x}}} - 1]$ in the unit interval by simply dividing $f(\mathbf{x})$ by $2^{l_{\mathbf{x}}} - 1$ and get our proposed numerical covariate z

$$z \;=\; g(x_1, \ldots, x_{l_{\mathbf{x}}}) = g(\mathbf{x}) = \frac{f(\mathbf{x})}{2^{l_{\mathbf{x}}} - 1} \in \left\{0, \frac{1}{2^{l_{\mathbf{x}}} - 1}, \frac{2}{2^{l_{\mathbf{x}}} - 1}, ..., 1\right\} \qquad (3.3)$$

From now on we will pretend that z is a continuous time-varying covariate. As a matter of fact the function $g(\mathbf{x})$ has a finite-discrete range. However, if we extend $\mathbf{x}$ to be a possibly infinite sequence we have that $\{g(\mathbf{x}) : \mathbf{x} \in H\}$ corresponds to the set of dyadic rationals in $[0, 1]$ which is a dense subset in $[0, 1]$.
At first sight this may be thought of only as a technical mathematical device, but it can have a plausibly realistic interpretation as a standardized quantization of the past experience or the accumulation of practice/training/memory with respect to the previously occurred events. In fact in the general setting of longitudinal binary outcomes one can consider binary events such as successful surgery experiences or correctly performed tasks and one can easily argue that the ordering induced by the quantization of the previous binary history is sensible. The same argument can be applied for a memory behavioural effect in a capture-recapture context.
Indeed the transformation $g(\mathbf{x})$ introduces a meaningful ordering of partial capture histories. In fact one can argue that in the process of learning from the past experience a 1 digit occurrence in the very last position of the binary string (last occasion) can affect the individual behaviour with a greater impact than a 1 digit in the previous occasions. Moreover, the more the 1 digits in the partial capture history the greater the impact. Of course we are not claiming the necessity of such ordering but we are explaining how it can be reasonably and fairly interpreted. Even though there is no compelling argument for the corresponding quantization it can be considered a convenient starting point to be refined further with alternative suitable data-driven rescaling such as the one in (3.3) or other transformations as in Section 3.3
Considering the function $g : H \to [0, 1]$ as in (3.3) and all the previous partial capture histories corresponding to the binary matrix $\mathbf{X} = [x_{ij}]$ one can derive a covariate matrix $\mathbf{Z} = [z_{ij}]$ as follows

$$z_{ij} = g(x_{i1}, \ldots, x_{ij-1}) \quad \forall i = 1, \ldots, N \; ; \; \forall j = 1, \ldots, t$$

Notice that the first column of $\mathbf{Z}$ corresponds to a null column since, for $j = 1$, the partial history $\mathbf{x} = (x_{i1}, \ldots x_{ij-1})$ corresponds in fact to an empty history ($\mathbf{x} = ()$).
We now show in practice how the covariate conversion works. We take as example

the same capture history analyzed in Chapter 2

$$(x_{i1}, \ldots, x_{i10}) \;=\; (0, 0, 1, 0, 0, 1, 1, 0, 0, 1)$$

We derive all the quantizations corresponding to all partial capture histories in Table 3.1

Time	**Current Occurrence**	**Partial capture history**	**Numeric Covariate**
j	x_{ij}	$(x_{i1}, \ldots, x_{ij-1})$	z_{ij}
1	0	( )	0.000
2	0	(0)	$0.000 = 0/1$
3	1	(0 , 0)	$0.000 = 0/3$
4	0	(0 , 0 , 1)	$0.571 = 4/7$
5	0	(0 , 0 , 1 , 0)	$0.267 = 4/15$
6	1	(0 , 0 , 1 , 0 , 0)	$0.129 = 4/31$
7	1	(0 , 0 , 1 , 0 , 0 , 1)	$0.571 = 36/63$
8	0	(0 , 0 , 1 , 0 , 0 , 1 , 1)	$0.787 = 100/127$
9	0	(0 , 0 , 1 , 0 , 0 , 1 , 1 , 0)	$0.392 = 100/255$
10	1	(0 , 0 , 1 , 0 , 0 , 1 , 1 , 0 , 0)	$0.196 = 100/511$

Table 3.1: Quantization of all partial capture histories corresponding to $\mathbf{x} = (0, 0, 1, 0, 0, 1, 1, 0, 0, 1)$

In our capture-recapture analysis we will use z_{ij} as an individual covariate changing with time j. For implementation purposes, both $\mathbf{X}$ and $\mathbf{Z}$ can be vectorized considering each double index i, j as a label for a single binary outcome x_{ij} whose probability can be explained in terms of the corresponding covariate z_{ij}. In the following we will start considering a simple linear logistic model as in (3.2) but other more flexible models can be adopted such as polynomial logistic regression, splines, etc. Notice that, differently from the usual covariates observable in a capture-recapture context from the sample stages (sex, age, length, etc) we do know the values of the z's also for the unobserved units. In fact, considering that units observed are labelled from 1 to M and those not observed are labelled from $M + 1$ to N we have

$$z_{ij} = 0 \quad \forall i = M + 1, \ldots N \; ; \; \forall j = 1, \ldots, t$$

We remark that other partial or total orderings can be considered sensible and useful in real data applications such as those based on the absolute or relative number of events experienced previously than time t. The reason why we are particularly interested in the ordering induced by $g(\mathbf{x})$ as in (3.3) is that it is a rather flexible

device which can be also used to reproduce Markov structure of arbitrary order. We will explain in detail the relationship between the continuous covariate z and the Markovian structure in Section 3.2. We illustrate alternative quantization of past experience in Section 3.3

Notice that considering a numeric covariate z built as described in (3.3) and a generic linear logistic regression model as in (3.2) the probability P_0 of never being captured during the whole experiment is

$$P_0 = \left(1 - \frac{e^{\alpha}}{1+e^{\alpha}}\right)^t$$

and depends only on the parameter α while β affects only the recapture probabilities. We notice that the probability P_0 depends only on one parameter (α) as in the class $\tilde{\mathcal{M}}$, but here we have a different structure linking capture probabilities corresponding to different partial capture histories and hence this alternative model cannot be recovered from the framework in Farcomeni (2011).

We hint also at a possible extension outside the closed capture-recapture context of the quantization idea. In fact it is possible to generalize this strategy to categorical-ordinal data using an appropriate scaling. Consider a generic ordinal variable with support $\{0, 1, \ldots, c\}$. Analogously to (3.3) we can quantify a generic sequence $(y_1, \ldots, y_{l_{\mathbf{x}}})$ where $y_j \in \{0, 1, \ldots, c\}$ as follows

$$z = \frac{\sum_{j=1}^{l_{\mathbf{x}}} y_j c^{j-1}}{c^{l_{\mathbf{x}}} - 1} \in \left\{0, \frac{1}{c^{l_{\mathbf{x}}} - 1}, \frac{2}{c^{l_{\mathbf{x}}} - 1}, ..., 1\right\}$$

3.2 Covariate representation and Markovian structure

In this subsection we go back to the topic of building behavioural models based on meaningful partitions of the subset H as in Chapter 2. We will show how the numeric covariate z can be also used to set up meaningful partitions of H and how one can go back to those partitions corresponding of Markovian models of order k. If we fix a positive integer $k < t$ we can partition the set H of all partial capture histories according to the value of $g(\mathbf{x})$ into appropriate subintervals namely

$$I_0 = \left[0, \frac{1}{2^k}\right], ..., I_{r-1} = \left(\frac{r-1}{2^k}, \frac{r}{2^k}\right], I_r = \left(\frac{r}{2^k}, \frac{r+1}{2^k}\right], ..., I_{2^k-1} = \left(\frac{2^k-1}{2^k}, 1\right] \quad (3.4)$$

Formally we get the partition $H = \{H_1, \ldots, H_{2^k}\}$ where

$$\mathbf{x} \in H_{r+1} \Leftrightarrow z = g(\mathbf{x}) \in I_r \quad \forall\, r \in \{0, 1, \ldots, 2^k - 1\} \quad (3.5)$$

so that the equivalence classes of binary subsequences depend only on the last k binary events. Hence, the mapping g defined in (3.3) is such that, for each partial capture history $\mathbf{x} \in H$, $z = g(\mathbf{x})$ belongs to the same set I_r according to the last k digits of the binary sequence. In order to prove this fact we can consider the following sum

$$\sum_{j=(l_{\mathbf{x}}-k+1)}^{l_{\mathbf{x}}} x_j \, 2^{j-1} \;=\; 2^{l_{\mathbf{x}}-k} \sum_{p=1}^{k} x_{l_{\mathbf{x}}-k+p} \, 2^{p-1}$$

which for $(x_{(l_{\mathbf{x}}-k+1)}, \dots x_{l_{\mathbf{x}}}) \in \mathcal{X}^k$ takes value in $2^{l_{\mathbf{x}}-k}\{0, 1, \dots, (2^k-1)\}$ and hence

$$\frac{\sum_{j=(l_{\mathbf{x}}-k+1)}^{l_{\mathbf{x}}} x_j \, 2^{j-1}}{2^{l_{\mathbf{x}}}-1} \;=\; \frac{\sum_{p=1}^{k} x_{l_{\mathbf{x}}-k+p} \, 2^{p-1}}{2^k - \frac{1}{2^{l_{\mathbf{x}}-k}}} \;\in\; \left\{0, \frac{1}{2^k - \frac{1}{2^{l_{\mathbf{x}}-k}}}, \dots, \frac{(2^k-1)}{2^k - \frac{1}{2^{l_{\mathbf{x}}-k}}}\right\} \tag{3.6}$$

On the other hand, considering the first $(l_{\mathbf{x}} - k)$ digits of the binary sequence we have

$$0 \;\leq\; \frac{\sum_{j=1}^{l_{\mathbf{x}}-k} x_j \, 2^{j-1}}{2^{l_{\mathbf{x}}}-1} \;\leq\; \frac{2^{l_{\mathbf{x}}-k}-1}{2^{l_{\mathbf{x}}}-1} \;=\; \frac{1 - \frac{1}{2^{l_{\mathbf{x}}-k}}}{2^k - \frac{1}{2^{l_{\mathbf{x}}-k}}} \tag{3.7}$$

For any $(x_{(l_{\mathbf{x}}-k+1)}, \dots x_{l_{\mathbf{x}}}) \in \mathcal{X}^k$ from (3.6) we can represent

$$\frac{\sum_{j=(l_{\mathbf{x}}-k+1)}^{l_{\mathbf{x}}} x_j \, 2^{j-1}}{2^{l_{\mathbf{x}}}-1} = \frac{r}{2^k - \frac{1}{2^{l_{\mathbf{x}}-k}}} \tag{3.8}$$

for some $r \in \{0, 1, \dots, (2^k-1)\}$ so that the following inequalities hold

$$\frac{r}{2^k} \;<\; \frac{r}{2^k - \frac{1}{2^{l_{\mathbf{x}}-k}}} \;\leq\; \frac{\sum_{j=1}^{l_{\mathbf{x}}} x_j \, 2^{j-1}}{2^{l_{\mathbf{x}}}-1} \;\leq\; \frac{r}{2^k - \frac{1}{2^{l_{\mathbf{x}}-k}}} + \frac{1 - \frac{1}{2^{l_{\mathbf{x}}-k}}}{2^k - \frac{1}{2^{l_{\mathbf{x}}-k}}} \;\leq\; \frac{r+1}{2^k} \tag{3.9}$$

Indeed the second inequality follows from the fact that the rhs has the sum running over all the elements of the binary sequence while the lhs corresponds to (3.8) where the sum runs over the last k elements only. The third inequality follows from (3.7). Finally, the last inequality follows from the fact that $\forall\, r \;\in\; \{0, 1, \dots, (2^k-1)\}$ we have

$$\frac{r+1-\frac{1}{2^{l_{\mathbf{x}}-k}}}{2^k - \frac{1}{2^{l_{\mathbf{x}}-k}}} - \frac{r+1}{2^k} = \frac{2^{l_{\mathbf{x}}-k}\left[(r+1) - \frac{1}{2^{l_{\mathbf{x}}-k}}\right]}{2^{l_{\mathbf{x}}}-1} - \frac{r+1}{2^k} = \frac{r+1-2^k}{(2^{l_{\mathbf{x}}}-1)2^k} \leq 0$$

In this way we have formally proved that for any $\mathbf{x}^{'}$ and $\mathbf{x}^{''}$ sharing the same last k digits we have that $g(\mathbf{x}^{'}) \in I_r$ and $g(\mathbf{x}^{''}) \in I_r$ for a suitable integer r such that

$$r = \frac{\sum_{j=(l_{\mathbf{x}}-k+1)}^{l_{\mathbf{x}}} x_j 2^{j-1}}{2^{l_{\mathbf{x}}-k}}$$

This implies that the elements H_r in the partition (3.5) are in correspondence with all binary configuration of the last k occurrences in each partial partial capture history. This will lead us back to the Markovian model of order k.
In order to get it straight we illustrate the following example where we show that no matter how the first $l_{\mathbf{x}} - k$ captures are arranged we get that $x = g(\mathbf{x})$ is always included in the same interval $I_r = I_2$. Consider fixed the last $k = 2$ digits. For all capture histories such that $(x_{l_{\mathbf{x}}-1}, x_{l_{\mathbf{x}}}) = (0, 1)$ we have

$$z = g(01) = \frac{0 \cdot 2^0 + 1 \cdot 2^1}{2^2 - 1} = 0.667 \in I_2 = \left(\frac{2}{2^2}, \frac{3}{2^2}\right]$$

$$z = g(001) = \frac{0 \cdot 2^0 + 0 \cdot 2^1 + 1 \cdot 2^2}{2^3 - 1} = 0.571 \in I_2 = \left(\frac{2}{2^2}, \frac{3}{2^2}\right]$$

$$z = g(101) = \frac{1 \cdot 2^0 + 0 \cdot 2^1 + 1 \cdot 2^2}{2^3 - 1} = 0.714 \in I_2 = \left(\frac{2}{2^2}, \frac{3}{2^2}\right]$$

$$z = g(0001) = \frac{0 \cdot 2^0 + 0 \cdot 2^1 + 0 \cdot 2^2 + 1 \cdot 2^3}{2^4 - 1} = 0.533 \in I_2 = \left(\frac{2}{2^2}, \frac{3}{2^2}\right]$$

$$z = g(1001) = \frac{1 \cdot 2^0 + 0 \cdot 2^1 + 0 \cdot 2^2 + 1 \cdot 2^3}{2^4 - 1} = 0.600 \in I_2 = \left(\frac{2}{2^2}, \frac{3}{2^2}\right]$$

$$z = g(0101) = \frac{0 \cdot 2^0 + 1 \cdot 2^1 + 0 \cdot 2^2 + 1 \cdot 2^3}{2^4 - 1} = 0.667 \in I_2 = \left(\frac{2}{2^2}, \frac{3}{2^2}\right]$$

$$z = g(1101) = \frac{1 \cdot 2^0 + 1 \cdot 2^1 + 0 \cdot 2^2 + 1 \cdot 2^3}{2^4 - 1} = 0.733 \in I_2 = \left(\frac{2}{2^2}, \frac{3}{2^2}\right]$$

and so on. The partition defined in (3.5) is equivalent to considering a general logistic regression as in (3.1) where the function $r(z)$ is a real step-function $s(z)$ which is constant over each subinterval I_r as follows

$$s(z) = \begin{cases} logit(p_{H_1}) & z \in I_0 = \left[0, \frac{1}{2^k}\right] \\ \dots & \\ logit(p_{H_r}) & z \in I_{r-1} = (\frac{r-1}{2^k}, \frac{r}{2^k}] \\ \dots & \\ logit(p_{H_{2^k}}) & z \in I_{2^k-1} = (\frac{2^k-1}{2^k}, 1] \end{cases}$$

where

$$p_{H_r} = P(X_{ij} = 1|(x_{i1}, \ldots, x_{ij-1}) \in H_r) = P(X_{ij} = 1|g(x_{i1}, \ldots, x_{ij-1}) \in I_{r-1})$$

according to the notation used in Chapter 2.
To see a first connection to the Markovian structure let us fix $t = 5$ and $k = 1$ and consider the following 2 subintervals which divide the unit interval representing the support of the variable z

$$I_0 = \left[0, \frac{1}{2}\right], I_1 = \left(\frac{1}{2}, 1\right]$$

From (3.5) the partition of the set H of all partial capture histories is the following partition for binary strings

$$\begin{cases} H_1 = \{(), (0), (00), (10), (000), (100), (010), (110), \\ \qquad (0000), (0100), (0010), (0110), (1000), (1100), (1010), (1110)\} \\ H_2 = \{(1), (01), (11), (001), (101), (011), (111), \\ \qquad (0001), (0011), (0101), (0111), (1001), (1011), (1101)(1111)\} \end{cases}$$

so that for any $\mathbf{x} \in H_r$, $r = 1, 2$ we get $g(\mathbf{x}) \in [(r-1)/2, r/2)$. The bipartition obtained is exactly the same as the one introduced in the previous section for model M_{c_1}.
Similarly, for $t = 2$ and $k = 2$ we have

$$I_0 = \left[0, \frac{1}{4}\right], I_1 = \left(\frac{1}{4}, \frac{2}{4}\right], I_2 = \left(\frac{1}{2}, \frac{3}{4}\right], I_3 = \left(\frac{3}{4}, 1\right]$$

and the partition of H corresponding to (3.5) is

$$\begin{cases} H_1 = \{(), (0), (00), (000), (100), (0000), (0100), (1000), (1100)\} \\ H_2 = \{(10), (010), (110), (0010), (0110), (1010), (1110)\} \\ H_3 = \{(01), (001), (101), (0001), (0101), (1001), (1101)\} \\ H_4 = \{(1), (11), (011), (111), (0011), (0111), (1011), (1111)\} \end{cases}$$

Notice however that, differently from the partition corresponding to model M_{c_2} considered in the previous chapter, the partial capture history (1) belongs to the subset H_4 instead of H_3. This in fact breaks the correspondence with the Markovian structure. To recover it we can make the following adjustment to the definition of the original numeric summary $g(\mathbf{x})$ slightly changing its argument $\mathbf{x}$. We augment the observed capture histories by conventionally imputing unobserved partial capture histories at the beginning. For the covariate construction only we assume that we

know that there are a suitable number $k-1$ of capture occasions preceding the first ones actually observed one and we pretend to know that they all resulted in no capture. Hence, if we insert a zero ahead on each actually observed partial capture history $(x_{i1},\dots,x_{ij-1})$ and denote it with $\mathbf{x}_{aug}$. We can then recover exactly the partition $\mathcal{H}_4(M_{c_2})$ as in Chapter 2 as follows

$$\begin{cases} H_1 = \{(\underline{0}), (\underline{0}0), (\underline{0}00), (\underline{0}000), (\underline{0}100), \\ \qquad (\underline{0}0000), (\underline{0}0100), (\underline{0}1000), (\underline{0}1100)\} \\ H_2 = \{(\underline{0}10), (\underline{0}010), (\underline{0}110), \\ \qquad (\underline{0}0010), (\underline{0}0110), (\underline{0}1010), (\underline{0}1110)\} \\ H_3 = \{(\underline{0}1), (\underline{0}01), (\underline{0}001), (\underline{0}101), \\ \qquad (\underline{0}0001), (\underline{0}0101), (\underline{0}1001), (\underline{0}1101)\} \\ H_4 = \{(\underline{0}11), (\underline{0}011), (\underline{0}111), \\ \qquad (\underline{0}0011), (\underline{0}0111), (\underline{0}1011), (\underline{0}1111)\} \end{cases}$$

which corresponds to

$$\mathbf{x}_{aug} \in H_r \;\Leftrightarrow\; g(\mathbf{x}_{aug}) \in I_{r-1}$$

Notice that we have marked the imputed initial segment with an underline sign. In this way the empty partial capture history () changes in $(\underline{0})$, (0) changes in $(\underline{0}0)$, (1) changes in $(\underline{0}1)$ and so on. We can turn this in matrix notation and consider the augmented binary capture history matrix. We denote with $\mathbf{X}_{aug} = [\underline{\mathbf{0}}, \mathbf{X}]$ the matrix obtained by adding a column of zeros on the left side of the matrix $\mathbf{X}$ and we can determine the corresponding covariate matrix $\mathbf{Z}_{aug}$ by applying the function g to all partial capture histories in $\mathbf{X}_{aug}$. At this point, instead of the former matrix $\mathbf{Z}$ built directly from $\mathbf{X}$ we use as covariate matrix corresponding to each entry of the original $\mathbf{X}$ only the last t columns of $\mathbf{Z}_{aug}$. The partition obtained in this way recovers exactly the partition $\mathcal{H}_4(M_{c_2})$. This example can be generalized to recover the partition structure of any model M_{c_k} of arbitrary Markov order k. In fact, to recover the partition $\mathcal{H}_2^k(M_{c_k})$ we need to consider the subintervals

$$\left(0, \frac{1}{2^k}\right], \left(\frac{1}{2^k}, \frac{2}{2^k}\right] \dots, \left(\frac{r-1}{2^k}, \frac{r}{2^k}\right], \dots, \left(\frac{2^k-1}{2^k}, 1\right]$$

and as covariate matrix the last t columns of $\mathbf{Z}_{aug}$ related to the augmented matrix $\mathbf{X}_{aug}$ obtained inserting $k-1$ columns of zeros on the left of the matrix $\mathbf{X}$.
It is easy to verify that the proposed adjustment is equivalent to considering the same subintervals $I_1,\dots,I_r,\dots,I_{2^k}$ for a numeric covariate z defined through a slightly modified g function denoted with g_M as follows

$$z = g_M(\mathbf{x}) = \sum_{j=1}^{l_{\mathbf{x}}} \frac{x_j 2^{[(j-1)+(k-1)]}}{2^{[(l_{\mathbf{x}}-1)+k]}-1} \in \left\{0, \frac{1}{2^{[(l_{\mathbf{x}}-1)+k]}-1}, \frac{2}{2^{[(l_{\mathbf{x}}-1)+k]}-1}, \dots, \frac{2^{[(l_{\mathbf{x}}-1)+(k-1)]}}{2^{[(l_{\mathbf{x}}-1)+k]}-1}\right\}$$

Notice that in this case the maximum value that $z = g_M(\mathbf{x})$ can take is no longer 1 but $\frac{2^{[(l_{\mathbf{x}}-1)+(k-1)]}}{2^{[(l_{\mathbf{x}}-1)+k]}-1}$.
To better illustrate how the adjusted procedure recovers the partition $\mathcal{H}_4(M_{c_2})$ consider the partial capture history (1). We have 2 different values for z according to whether or not we add the auxiliary zero ahead of the capture history.

$$(1) \;\Rightarrow\; z = g(1) = \frac{1 \cdot 2^0}{2^1 - 1} = 1$$

$$(\underline{0}, 1) \;\Rightarrow\; z = g(0,1) = g_M(1) = \frac{0 \cdot 2^0 + 1 \cdot 2^1}{2^2 - 1} = 0.57$$

Only with the modified function g_M the partial capture history $\mathbf{x} = (1)$ yields the appropriate z-value which makes $\mathbf{x} = (1)$ belong to the correct equivalence class H_3. In general both partitions induced by $g(\mathbf{x})$ and $g_M(\mathbf{x})$ with values in the previous subintervals I_r are reasonable options but only the second one recovers exactly the Markovian structure M_{c_k}. Notice also that the ordering induced by g_M allows for example to distinguish between (1) and (1,1,1,1,1) while that is not possible using (3.3).

We now sketch a list of other meaningful alternatives for partitioning the covariate range. Indeed, it is possible to recover model M_b associated to the partition $\mathcal{H}_2(M_b)$ by partitioning the support of $z = g(\mathbf{x})$ as follows

$$I_1 = \left[0, \frac{1}{2^t}\right], I_2 = \left(\frac{1}{2^t}, 1\right]$$

so that the regression step function becomes

$$s(z) = \begin{cases} logit(p) & z \in I_1 \\ \\ logit(r) & z \in I_2 \end{cases}$$

Notice that the first partition I_1 can be equivalently reduced to the single value $\{0\}$. In general each model corresponding to the partition of the range of z i.e. the unit interval of the form $\{I_1, I_2 \ldots I_B\} = \{[0, e_1], (e_1, e_2], \ldots, (e_{B-1}, 1]\}$ for $r \geq 1$, represents a model $M^\star$ associated to the partition $\mathcal{H}_B(M^\star)$ such that $M^\star$ belongs to the class $\tilde{\mathcal{M}}$ formalized in Chapter 2. Formally model $M^\star$ is equivalent to (3.1)

with the following step-function

$$s(z) = \begin{cases} logit(p_{H_1}) & z \in [0, e_1] \\ \dots & \\ logit(p_{H_b}) & z \in (e_{b-1}, e_b] \\ \dots & \\ logit(p_{H_B}) & z \in (e_{B-1}, 1] \end{cases}$$

This representation embeds some of the original models proposed in Farcomeni (2011) such as M_L. In fact, we notice that model M_{L_2} can be recovered within our general logistic regression (3.1) which relies on the meaningful numeric covariate z adopting as a regression function one which is a step function with only two levels corresponding to the bipartition of the range of z into two contiguous intervals: $[0, 0.625]; (0.625, 1]$. In fact, only for partial capture histories (1), (01) and for all $\mathbf{x}$ with $l_{\mathbf{x}} \geq 3$ such that $(x_{i1}, \dots, x_{il_{\mathbf{x}-3}}, 0, 1, 1)$, $(x_{i1}, \dots, x_{il_{\mathbf{x}-3}}, 1, 0, 1)$ and $(x_{i1}, \dots, x_{il_{\mathbf{x}-3}}, 1, 1, 1)$ we have that $z > 0.625$. Hence these intervals lead to the same bi-partition of the set H considered in the model M_{L_2}.

3.3 Alternative meaningful numerical behavioural covariates

As previously highlighted, the procedure of ordering and scaling a generic partial capture history defined in (3.3) is not the only way of representing the quantization of a binary sequence. Indeed, although we have argued how that choice can be considered reasonable in some cases (also in terms of Markovian structure) it can be open for some criticism. For example, consider the following two partial capture histories each based on five capture occasions

$$\mathbf{x}_1 = (1, 1, 1, 1, 0) \quad \Rightarrow \quad g(\mathbf{x}_1) = \frac{1 \cdot 2^0 + 1 \cdot 2^1 + 1 \cdot 2^2 + 1 \cdot 2^3 + 0 \cdot 2^4}{2^5 - 1} = \frac{15}{31} = 0.484$$

$$\mathbf{x}_2 = (0, 0, 0, 0, 1) \quad \Rightarrow \quad g(\mathbf{x}_2) \frac{0 \cdot 2^0 + 0 \cdot 2^1 + 0 \cdot 2^2 + 0 \cdot 2^3 + 1 \cdot 2^4}{2^5 - 1} = \frac{16}{31} = 0.516$$

The first partial capture history $\mathbf{x}_1$ is characterized by having a total of four captures in the first four occasions while the second one $\mathbf{x}_2$ has only one capture in the last occasion. The mapping described in (3.3) assigns a larger impact on the conditional probabilities to $\mathbf{x}_2$. One can find undesirable the fact that a partial capture history having just a single capture, even though in the last occasion, gives rise to a larger value compared to a binary sequence which has 4 captures out of 5.

As a possible alternative useful mapping one can consider a function based on the total number of captures occurred for each partial capture history $\mathbf{x} \in H$. In order to obtain a potentially continuous covariate as in (3.3) we rescale the range in the unit interval considering as denominator the length of each capture history as follows

$$z = g_n(\mathbf{x}) = g_n(x_1, \ldots, x_{l_\mathbf{x}}) = \frac{\sum_{j=1}^{l_\mathbf{x}} x_j}{l_\mathbf{x}} \in \left\{0, \frac{1}{l_\mathbf{x}}, \frac{2}{l_\mathbf{x}}, ..., 1\right\} \tag{3.10}$$

The partial capture histories $\mathbf{x}_1$ and $\mathbf{x}_2$ can be quantified as in (3.10)

$$\begin{aligned} \mathbf{x}_1 = (1,1,1,1,0) \quad &\Rightarrow \quad \frac{4}{5} = 0.8 \\ \mathbf{x}_2 = (0,0,0,0,1) \quad &\Rightarrow \quad \frac{1}{5} = 0.2 \end{aligned}$$

It is also possible to rescale the number of captures by considering the total number of occasions in the whole experiment as follows

$$z = \tilde{g}_n(\mathbf{x}) = \tilde{g}_n(x_1, \ldots, x_{l_\mathbf{x}}) = \frac{\sum_{j=1}^{l_\mathbf{x}} x_j}{t} \in \left\{0, \frac{1}{t-1}, \frac{2}{t-1}, ..., 1\right\} \tag{3.11}$$

On the other hand the mapping g_n and $\tilde{g}_n$ described in (3.10) and (3.11) may have in turn their own undesirable features. In fact they do not take into account the inner sequence structure considering the number of captures only. For example a partial capture history (1,0,0,0,0) will be equivalent in terms of g_n and $\tilde{g}_n$ to $\mathbf{x}_2$ even though they are substantially different.

3.4 Unconditional maximum likelihood inference

In this section we will show a simple-to-implement procedure to infer on the parameter space through the unconditional likelihood which yields as by product inference on the main parameter of interest N using the profile likelihood. Indeed, the new approach exploiting a numerical summary of partial capture histories and logistic regression framework allows to recycle consolidated standard GLM routines in our capture-recapture context.

Let N_{upp} be a suitably high fixed upperbound for the population size. In order to make inference on N one has to evaluate, for each fixed value of $N \in \{M, M+1, \ldots, N_{upp}\}$, the maximum of the unconditional likelihood function (UMLE) obtained for a standard logistic model fitted using $N \times t$ binary observations with the corresponding numerical covariates z_{ij}. Let $L(N, \alpha, \beta)$ be the likelihood function for the linear logistic model (3.1) such that

$$L(N,\alpha,\beta) \propto \binom{N}{M} \left[\prod_{i=1}^{N}\prod_{j=1}^{t} \left(\frac{\exp(\alpha+\beta z_{ij})}{1+\exp(\alpha+\beta z_{ij})}\right)^{x_{ij}} \left(1 - \frac{\exp(\alpha+\beta z_{ij})}{1+\exp(\alpha+\beta z_{ij})}\right)^{1-x_{ij}}\right] \tag{3.12}$$

and denote with

$$\hat{L}(N) = L(N, \hat{\alpha}(N), \hat{\beta}(N))$$

the resulting value of the maximized likelihood. One can get the maximum likelihood estimate for N by maximizing such profile likelihood $\hat{L}(N)$ obtained for each fixed N with a standard GLM routine. Unconditional maximum likelihood estimate for N will then be

$$\hat{N} = \arg \max_{N \in \{M, \ldots, N_{upp}\}} \left(\hat{L}(N) \right)$$

Notice that, given $\hat{N}$, the joint unconditional likelihood for all parameters involved in the model is globally maximized as $L(\hat{N}, \hat{\alpha}(\hat{N}), \hat{\beta}(\hat{N}))$. The estimate procedure requires to iteratively fit a logistic regression for each $N \in \{M, \ldots, N_{upp}\}$. For large values of N_{upp} this procedure can be computationally demanding and time-consuming involving logistic procedures repeated $N_{upp} - M + 1$ times. To reduce computational effort and computing time it is possible to evaluate the profile likelihood function not for each values of N but only on a suitable sub grid and use some parallel computing environment to run simultaneously multiple logistic fits.
In Chapter 2 we have seen that using the (complete) unconditional likelihood the likelihood failure problem is not overcome for models belonging to the class $\tilde{\mathcal{M}}$. In this case we are outside this model framework and we could not find suitable theoretical conditions for its possible occurrence. We address likelihood failure occurrence only on the basis of simulated data in Section 3.6. Although we did not encounter likelihood failure in our simulation study we experienced sometimes a nearly flat likelihood. As justified in Chapter 2 we propose a Bayesian approach to address this issue.

3.5 Alternative implementation of Bayesian Inference

Motivated from the simulation results of Chapter 2 we are interested in exploring the performance of a Bayesian analysis even if the likelihood failure pathology does not occur.
Since in this case the model structure does not allow to derive integral quantities in closed form we use MCMC simulations. We propose an implementation based on a Metropolis-within-Gibbs (MWG) algorithm where the simulations from the full-conditionals are replaced by simulations from a Metropolis-Hastings kernel having the full-conditional as its target distribution

$$\pi_{ARMS}(\theta_i | \theta_{-i}, data)$$

where $\theta = (N, \alpha, \beta)$. Let R be the total number of iterations. The MWG algorithm is schematically described in (3.13) and is implemented using the function `arms` of the `R` package `HI` which allows to perform Adaptive Rejection Metropolis Sampling (ARMS)

$$\begin{aligned} &Step\ 0 : N^{(0)} = M\ ;\ \alpha^{(0)} = 0\ ;\ \beta^{(0)} = 0 \\ \\ &Step\ r : N^r \sim \pi_{ARMS}(N|\alpha^{r-1}, \beta^{r-1}, data) \qquad r = 1, \ldots, R \\ &\qquad\quad \alpha^r \sim \pi_{ARMS}(\alpha|N^r, \beta^{r-1}, data) \\ &\qquad\quad \beta^r \sim \pi_{ARMS}(\beta|N^r, \alpha^r, data) \end{aligned} \tag{3.13}$$

where for a generic step only one point is sampled from the target distribution. Notice that N is considered in our implementation as a continuous parameter. More rigorously, we could sample from the discrete full-conditional distribution for N

$$\pi(N|\alpha^{(r-1)}, \beta^{(r-1)}) \propto \pi(N)L(N, \alpha^{(r-1)}, \beta^{(r-1)}) \qquad ;\ N = M, \ldots, N_{upper}$$

However we have verified (empirically) that this change does not lead to different results and hence we preferred to fully exploit the `arms` function which turns out to be computationally less expensive. In order to make the Bayesian implementation easier we decided to implement also another approximate simulation-based Bayesian inference in a capture-recapture context with the approach proposed in Royle et al. (2007). In their paper it is presented an alternative reparameterization of a discrete-time closed capture-recapture model based on a data-augmentation scheme allowing for an easy implementation of the Bayesian inference through the use of general purpose Bayesian modelling software such as `WinBUGS` or `JAGS`. The approach proposed in Royle et al. (2007) relies on a super-population structure borrowed from models for occupancy.

The basic idea is to bound the range of the possible number of unseen units with $N_{upp} - M$ and use an alternative data augmentation to reformulate the problem into one in which there is a super-population of fixed size N_{upp} from which the actual unknown population size N is randomly derived introducing an extra layer of latent structure. This can been seen as a convenient reparameterization of a discrete-time closed capture-recapture model.

More precisely, let N_{upp} be the size of the super-population which is interpreted as a known upperbound for the actual population size N. The procedure consists of augmenting the observed data set of size M with a known number $N_{upp} - M$ of rows of all zeros capture histories which can be considered as the unobserved histories of all "pseudo" units.

Besides the whole $N_{upp} \times t$ binary matrix $\mathbf{X}_{[N_{upp} \times t]}$ of capture histories there are N_{upp} new binary labels denoted with v_i which are used to distinguish which units of the super-population are part of the actual (partially observed) target population. More formally, consider the latent binary variable v_i representing for each $i = 1, \dots N_{upp}$, whether the i-th row associated to a unit of the super-population belongs to the target population or not

$$v_i = \begin{cases} 1 & \text{if unit } i \text{ of the super-population belongs to} \\ & \text{the actual partially observed population} \\ 0 & \text{otherwise} \end{cases}$$

The data augmentation structure is shown in Table 3.2. Of course the target unknown population size N corresponds to

$$N = \sum_{i=1}^{N_{upp}} v_i$$

Notice that if unit i is observed during the trapping stages it is certain that this

Unit	1	2	·	j-1	j	j+1	·	t-1	t	v_i
1	1	1	⋯	1	1	1	⋯	1	1	1
2	1	0	⋯	0	1	0	⋯	1	0	1
3	0	1	⋯	1	1	1	⋯	1	1	1
4	0	0	⋯	0	0	0	⋯	1	1	1
5	1	0	⋯	1	1	0	⋯	0	1	1
6	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	1
M	0	0	⋯	0	1	0	⋯	1	1	1
$M+1$	0	0	⋯	0	0	0	⋯	0	0	?
$M+2$	0	0	⋯	0	0	0	⋯	0	0	?
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	?
N_{upp}	0	0	⋯	0	0	0	⋯	0	0	?

Table 3.2: Data augmentation structure

unit belongs to the target population. Hence the conditional probability that the corresponding pseudo-unit belongs to the target population is equal to 1 and the corresponding latent variable v_i is equal to 1. On the other hand if $\sum_{j=1}^{t} x_{ij} = 0$ there is a non trivial chance that the unit is in fact part of the target population.

The inclusion probability structure can be formalized as follows

$$Pr(V_i = 1|\sum_{j=1}^{t} x_{ij}) = \begin{cases} 1 & \text{if } \sum_{j=1}^{t} x_{ij} \geq 1 \\ \psi_c = \frac{P_0\psi}{1-\psi(1-P_0)} & \text{otherwise} \end{cases}$$

In fact

$$Pr(V_i = 1|\sum_{j=1}^{t} x_{ij}) = \frac{Pr(V_i = 1 \cap \sum_{j=1}^{t} x_{ij})}{Pr(\sum_{j=1}^{t} x_{ij})} =$$
$$\frac{Pr(\sum_{j=1}^{t} x_{ij}|V_i = 1)Pr(V_i = 1)}{Pr(\sum_{j=1}^{t} x_{ij}|V_i = 1)Pr(V_i = 1) + Pr(\sum_{j=1}^{t} x_{ij}|V_i = 0)Pr(V_i = 0)} \Rightarrow$$

$$\begin{cases} Pr(V_i = 1|\sum_{j=1}^{t} x_{ij} \geq 1) = \frac{(1-P_0)\cdot\psi}{(1-P_0)\cdot\psi+0\cdot(1-\psi)} = 1 \\ Pr(V_i = 1|\sum_{j=1}^{t} x_{ij} = 0) = \frac{P_0\cdot\psi}{P_0\cdot\psi+1\cdot(1-\psi)} = \frac{P_0\psi}{1-\psi(1-P_0)} = \psi_c \end{cases}$$

and hence we have

$$\psi = Pr(V_i = 1) = \frac{P_0\psi_c}{\psi_c + P_0(1-\psi_c)}$$

where P_0 is the probability of never being observed during all trapping stages as defined in the previous chapter. On the other hand if a unit does not belong to the target population the probability of being captured is zero

$$Pr(x_{ij} = 1|v_i = 0) = 0 \qquad \forall j$$

The hierarchical super-population model can be summarized as follows

$$\begin{cases} [x_{ij}|v_i] \sim Bern\left(\left[\frac{\exp(\alpha+\beta z_{ij})}{1+\exp(\alpha+\beta z_{ij})}\right]^{v_i}\right) \\ \\ v_i \sim Ber(\psi) \qquad i = 1, \ldots, N_{upp} \end{cases} \tag{3.14}$$

From (3.14) it follows that

$$N|N_{upp} = \sum_{i=1}^{N_{upp}} v_i \sim Bin(N_{upp}, \psi)$$

So that, conditionally on N this super-population structure implies that $M = \sum_{i=1}^{N} I(\sum_{j=1}^{t} X_{ij} > 0) \sim Bin(N, 1-P_0)$ as in the original discrete-time closed population model. Under this formulation N is a derived parameter and the objective of the analysis is moved towards the latent parameter ψ. Hence, the new inference problem is to partition the pseudo-individuals into two groups: units belonging to the population ($v_i = 1$) and units which do not belong to the population

$(v_i = 0)$.

In implementing a Bayesian approach to the analysis of the linear regression model (3.1) , we require prior distributions for the model parameters N, α and β. However, under data augmentation, the distribution of the parameter of interest $N = \sum_{i=1}^{N_{upp}} v_i$ is implied by the distribution of the parameter ψ representing the probability of belonging to the actual target population. If one elicits a fixed value ψ_0 of ψ one obtains automatically a prior distribution on N such that

$$N|N_{upp} \sim Bin(N_{upp}, \psi_0)$$

This may be a too restricted shape for the prior on N. Hence in Royle et al. (2007) it is explained that eliciting a uniform hyper-prior on ψ leads indeed to a discrete uniform (DU) prior for N (conditioning on N_{upp})

$$\int_0^1 Bin(N_{upp}, \psi) U_\psi(0,1) d\psi = \int_0^1 \binom{N_{upp}}{N} \psi^N (1-\psi)^{N_{upp}-N} d\psi =$$
$$\binom{N_{upp}}{N} B(N+1, N_{upp}-N+1) = \frac{1}{N_{upp}+1}$$

$$\Rightarrow \quad N|N_{upp} \sim DU\{0, 1, \ldots, N_{upp}\}$$

We can easily extend the original idea proposed by Royle et al. (2007) choosing the hyper-prior for ψ within the beta family

$$\int_0^1 Bin(N_{upp}, \psi) Beta_\psi(a,b) d\psi = \tag{3.15}$$
$$\int_0^1 \left[\binom{N_{upp}}{N} \psi^N (1-\psi)^{N_{upp}-N} \right] \times \left[\frac{1}{Beta(\alpha,\beta)} \psi^{\alpha-1} (1-\psi)^{\beta-1} \right] d\psi$$
$$\binom{N_{upp}}{N} \frac{1}{Beta(\alpha,\beta)} \int_0^1 \psi^{N+\alpha-1} (1-\psi)^{N_{upp}-N+\beta-1} d\psi$$

obtaining a Beta-Binomial distribution

$$\pi(N|N_{upp}, \alpha, \beta) = \binom{N_{upp}}{N} \frac{Beta(N+\alpha, N_{upp}-N+\beta)}{Beta(\alpha,\beta)} =$$

$$\frac{\Gamma(N_{upp}+1)}{\Gamma(N+1)\Gamma(N_{upp}-N+1)} \frac{\Gamma(N+\alpha)\Gamma(N_{upp}-N+\beta)}{\Gamma(N_{upp}+\alpha+\beta)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

Motivated from the simulation results in Chapter 2 we propose alternative prior distributions on ψ in order to obtain prior distributions for $N|N_{upp}$ such as $\pi(N) \propto 1/N$ and $\pi(N) \propto 1/N^2$. Prior $1/N$ can be obtained considering as prior distribution for ψ a Beta distribution with parameters $a \to 0$ and $b = 1$, in fact, ignoring the constant terms

$$\lim_{\alpha \to 0} \pi(N|N_{upp}, \alpha, \beta = 1) \propto \frac{\Gamma(N)}{\Gamma(N+1)} = \frac{1}{N}$$

Notice that, although the Beta distribution is parametrized by two positive parameters ($\alpha > 0$ and $\beta > 0$), the integral

$$\int_0^1 \binom{N_{upp}}{N} \psi^{N+\alpha-1}(1-\psi)^{N_{upp}-N+\beta-1} d\psi$$

can be well defined up to a proportionality constant also for $\alpha \leq 0$. Hence we can implement the prior $\pi(N) \propto 1/N^2$ considering $\alpha = -1$ and $\beta = 1$ in (3.16) such that

$$\pi(N|N_{upp}, \alpha = -1, \beta = 1) \propto \frac{\Gamma(N)\Gamma(N_{upp}+1)}{\Gamma(N+1)\Gamma(N_{upp})} = \frac{1}{N^2}$$

The implementation of this specific prior can be actually implemented in `JAGS` using a truncated Pareto distribution on ψ

$$\psi \sim Par(1, 0.0001)I(0.0001, 1)$$

We consider two different non-informative prior distributions for the nuisance parameters α and β consisting of a flat normal distribution

$$\pi_f(\alpha) = \pi_f(\beta) = N(0, 10^6)$$

and suitable hand-tuned independent prior distributions on α and β such that one obtains an almost flat distribution for the number of distinct units observed during whole experiment M. In order to achieve this goal in implementing our analysis we have considered

$$\pi_{U(M)}(\alpha) = N(-2, 1.4) \quad ; \quad \pi_{U(M)}(\beta) = N(-3, 3)$$

We point out that this implementation does not provide posterior distributions from the original model described in (3.2). To our knowledge this fact has not been previously noticed in similar standard model such as M_b. In fact the actual model implemented in `JAGS` corresponds to a slightly different likelihood structure sharing the same building blocks with a logistic regression with respect to the numeric covariate z_i only for those units of the super-population such that $v_i = 1$. The likelihood function is

$$\begin{aligned} L(\psi, \alpha, \beta) = & \binom{N_{upp}}{\sum_{i=1}^{N_{upp}} v_i} \psi^{\sum_{i=1}^{N_{upp}} v_i} (1-\psi)^{N_{upp} - \sum_{i=1}^{N_{upp}} v_i} \\ & \prod_{i=1}^{N_{upp}} \prod_{j=1}^{t} \left[\left(\frac{\exp(\alpha + \beta z_{ij})}{1 + \exp(\alpha + \beta z_{ij})} \right)^{x_{ij}} \left(1 - \frac{\exp(\alpha + \beta z_{ij})}{1 + \exp(\alpha + \beta z_{ij})} \right)^{1 - x_{ij}} \right]^{v_i} \end{aligned}$$

where it is easy to understand how $L(\psi, \alpha, \beta)$ depends on the upperbound N_{upp} and hence the likelihood structure is different from (3.12).

3.6 Simulation study

To verify the comparative performance of Bayesian versus classical unconditional likelihood approach we consider a simple simulation study where data are generated from a linear logistic regression (3.2) with two different configurations for the parameters α and β. The true value of the target population is $N = 100$. For each configuration $K = 100$ data-sets are generated considering $t = 5$ capture occasions. Alternative parameter configurations generating simulated data are described in Table 3.6 as trial Tr.1 and Tr.2

Trial	**Model**	**Parameters**	$E[M]/N$
$Tr.1$	$\text{logit}\left(p_j(x_1, \ldots x_{j-1})\right) = \alpha + \beta z_{ij}$	$\alpha = -log(4); \beta = 1.5$	0.67
$Tr.2$	$\text{logit}\left(p_j(x_1, \ldots x_{j-1})\right) = \alpha + \beta z_{ij}$	$\alpha = -log(9); \beta = 2$	0.41

Table 3.3: *Parameter configurations for simulation experiments.*

As noticed above the probability P_0 depends only on α. Hence to set up the two different configurations we selected two different values for α determining the same values of P_0, and correspondingly the same expected values of M, considered in the simulation study in the previous chapter (see Table 2.4). To summarize the posterior distribution of the population size N we consider the mode and the loss minimizer m_R for the loss function described in the previous chapter. Moreover, as interval estimates we have considered both equal tail and HPD intervals. In Table 3.4 and 3.6 we show the performance of both classical and Bayesian approaches respectively. Although the likelihood failure never occurred in the present simulation, we get overall results which are qualitatively similar to those obtained in the previous simulation study. The Bayesian approach outperforms the unconditional likelihood in terms of both point and interval estimates: better empirical RMSE and shortest interval estimates with actual empirical coverage sufficiently close to the nominal level 0.95.

Tr.	RMSE	coverage %	CI.length
1	0.69	96.2	285.25
2	0.17	95.4	79.34

Table 3.4: Simulation results: classical approach

As we can see the results from our Metropolis within Gibbs implementation shown

Tr.	Prior	$RMSE_{mode}$	$RMSE_{m_R}$	$\%_{HPD}$	l_{HPD}
1	U(N)-flat(α, β)	0.664	0.471	96.0	356.5
	$1/N$-flat(α, β)	0.395	0.325	97.0	251.68
	Rissanen-flat(α, β)	0.416	0.299	97.0	221.69
	$1/N^2$-flat(α, β)	0.309	0.272	95.0	178.95
	U(N)-Ucov(α, β)	0.285	0.253	99.0	150.73
	$1/N$-Ucov(α, β)	0.297	0.247	97.0	128.02
	Rissanen-Ucov(α, β)	0.289	0.250	96.0	119.95
	$1/N^2$-Ucov(α, β)	0.295	0.261	95.0	110.51
2	U(N)-flat(α, β)	0.195	0.192	98.0	92.79
	$1/N$-flat(α, β)	0.169	0.159	98.0	79.55
	Rissanen-flat(α, β)	0.154	0.151	99.0	75.69
	$1/N^2$-flat(α, β)	0.141	0.140	98.0	68.62
	U(N)-Ucov(α, β)	0.160	0.158	99.0	75.34
	$1/N$-Ucov(α, β)	0.135	0.141	98.0	69.98
	Rissanen-Ucov(α, β)	0.142	0.138	98.0	67.85
	$1/N^2$-Ucov(α, β)	0.147	0.131	98.0	64.59

Table 3.5: Simulation results: Bayesian analysis using `arms`.

in Table 3.5 and the super-population approach developed in `JAGS` shown in Table 3.6 are very close in terms of both RMSE and interval coverage. Moreover, it is also shown how different combinations of $\pi(N)$, prior distributions for the nuisance parameters $\pi(\alpha)$ and $\pi(\beta)$ and summary statistics of the posterior distribution for N yield a better performance when one chooses m_R as posterior summary, $\pi(N) \propto 1/N^2$ and the ad-hoc prior distribution on the nuisance parameters which yields an almost flat distribution of M (see Figure 3.1). In this case, slightly better than what has been experienced in the previous simulation study, we have a good frequentist coverage also when we consider $1/N^2$ as prior distribution for N.

3.7 Great Copper data

In this section we reanalyze Great Copper data-set already seen in Chapter 2. Data consist in a sample of 45 butterflies observed in 8 trapping occasions. We have implemented both classical and Bayesian inference as described in the previous sections.

Tr.	Prior	$RMSE_{mode}$	$RMSE_{m_R}$	$\%_{HPD}$	l_{HPD}
1	U(N)-flat(α, β)	1.22	0.67	95.0	416.60
	$1/N$-flat(α, β)	0.43	0.34	98.0	284.55
	$1/N^2$-flat(α, β)	0.32	0.27	93.0	172.64
	U(N)-Ucov(α, β)	0.32	0.26	98.0	154.36
	$1/N$-Ucov(α, β)	0.29	0.25	98.0	127.08
	$1/N^2$-Ucov(α, β)	0.29	0.26	94.0	109.96
2	U(N)-flat(α, β)	0.17	0.20	97.0	92.35
	$1/N$-flat(α, β)	0.15	0.16	98.0	80.31
	$1/N^2$-flat(α, β)	0.14	0.14	98.0	69.45
	U(N)-Ucov(α, β)	0.16	0.16	99.0	75.52
	$1/N$-Ucov(α, β)	0.15	0.14	99.0	69.29
	$1/N^2$-Ucov(α, β)	0.14	0.13	98.0	63.14

Table 3.6: Simulation results: Bayesian approach using `JAGS`.

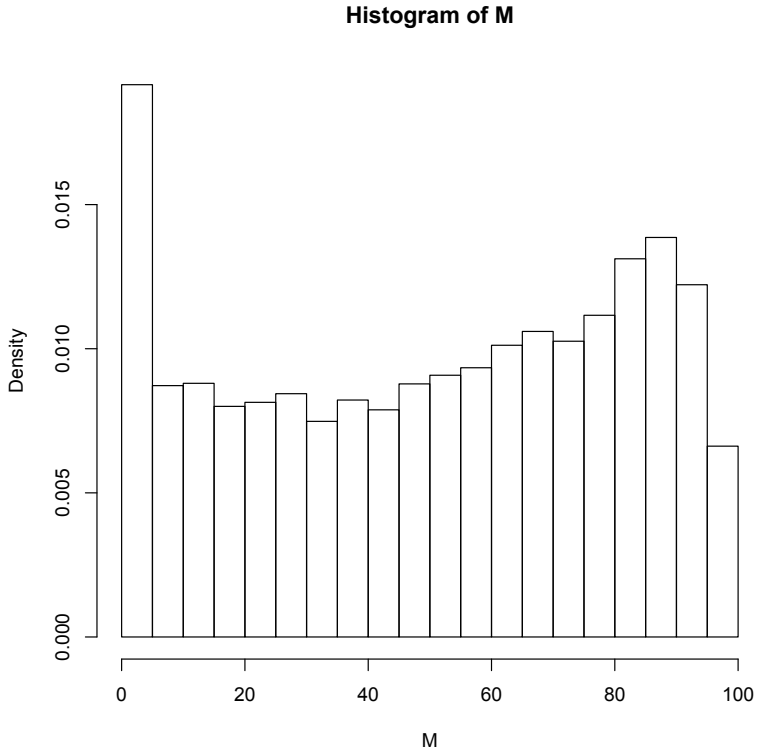

Figure 3.1: Empirical sample coverage obtained with an ad-hoc prior distributions on α

Model	**Approach**	**# parameters**	$\hat{N}$	(N^-, N^+)
Linear logistic	Classic: UMLE	2+1	170	(86,449)

Table 3.7: Great Copper data: Linear logistic model estimates via UMLE

Model	**Approach**	**# parameters**	$\hat{N}$	(N^-, N^+)
	Bayes: $1/N$ - flat	2+1	155	(70,427)
	Bayes: $1/N^2$ - flat		135	(70,352)
	Bayes: $1/N$ - U_M		117	(73,205)
	Bayes: $1/N^2$ - U_M		109	(70,184)

Table 3.8: Great Copper data: Linear logistic model estimates via `JAGS`

As we can see, from Table 3.7, 3.8 and 3.9 our linear logistic model yields re-

Model	**Approach**	**# parameters**	$\hat{N}$	(N^-, N^+)	log-ML
	Bayes: $1/N$ - flat	2+1	147	(71,403)	-164.76
	Bayes: Rissanen - flat		143	(69,368)	-164.63
	Bayes: $1/N^2$ - flat		132	(69,303)	-164.85
	Bayes: $1/N$ - U_M		144	(72,367)	-166.29
	Bayes: Rissanen - U_M		136	(67,335)	-166.22
	Bayes: $1/N^2$ - U_M		131	(67,318)	-165.93

Table 3.9: Great Copper data: Linear logistic model estimates, AIC and log marginal likelihood via `arms`

sults which represent a compromise between models involving the enduring effect only and the ephemeral effect model M_{c_k}. Moreover, Bayesian inference via super-population approach, differently from the one developed from the original model is very sensitive to the prior choice leading to somewhat different results especially in terms of widths of the interval estimates.

In Chapter 2 we have shown that AIC index and the log marginal likelihood (log-ML) both supported the one denoted in Farcomeni (2011) with M_{L_2} as the best model. The AIC index associated to the proposed linear logistic model is 319.46 and compared to the AIC index of the model considered in the previous chapter makes this new model as the best fitting.

In Bayesian analyses for model selection purpose we need to compute the log marginal likelihood. This is not a trivial task with our non-conjugate model. We propose a marginal likelihood estimation via power posteriors as suggested in Friel and Petit

(2008). The power posterior technique is not easily implementable in `JAGS` and so we only consider its estimation for our Metropolis within Gibbs implementation of the original model. As we can see in Table 3.9 the log marginal likelihood agrees with the AIC index suggesting model (3.2) as the best one among all candidate models.

3.8 Final remarks

In order to handle the behavioural effect to capture we have proposed an alternative flexible model framework based on a suitable ordering and scaling of the binary sequences representing the individual partial capture histories h. The proposed ordering consists of considering the sequence of any partial capture history as the binary representation of an integer. Then, in order to obtain a suitable quantitative covariate $z \in [0, 1]$, representing a numerical quantification of the partial capture history, the integer quantity is appropriately rescaled.
We provide some natural interpretation of the covariate z as a meaningful proxy for a *memory effect* and discuss some other alternative quantifications. The basic idea of the new model framework is to set-up a linear logistic model where each capture occurrence x_{ij} is considered as a binary outcome and $\alpha + \beta z_{ij}$ or, more generally, $r(z_{ij})$ as the linear predictor of the log-odds of the corresponding probability. Moreover, we pointed out how our meaningful numeric covariates z allows to recover classical behavioural models ($M_b, M_{c_k}, M_{c_k b}$ etc.) and many others (e.g. model M_L) by using a more flexible non-linear logistic regression in terms of an appropriate real step-function $s(z)$.
We implemented both classical inference recycling consolidated standard GLM routines and Bayesian inference in two alternative software implementations: one recycling simple-to-implement `BUGS-JAGS` scripting and a customized MCMC code written in R. As in Chapter 2 we investigated the sensitivity of the analysis with respect to few alternative default priors for the population size N and for the nuisance parameters α and β using, once again, their frequentist properties as performance criterion. We get a better performance when one chooses a prior distribution for N proportional to $(1/N^2)$ and two independent ad-hoc prior distributions on the nuisance parameters yielding an almost flat distribution of M. Although likelihood failure did not occur in our simulation study, it shown how, occasionally, a critical flat likelihood still persists also in this set-up. We have seen how Bayesian analysis mitigates the likelihood flatness problem reducing the relative mean square error and leading to shorter interval estimates in the presence of the same frequentist

coverage as occurred in Chapter 2.

Part III

Heterogeneity Effect Modeling

Chapter 4

Bayesian mixtures of Poisson modeling for capture-recapture experiments

Count data is increasingly common in data analysis. In fact many discrete responses have count data as possible outcome. In many scientific disciplines such as capture-recapture experiments, species richness problems, genomic applications, etc. data can be often expressed as a series of counts. Indeed, count data often represents the number of occurrences in a fixed period of time: in capture-recapture setting it can be the number of captures occurred during whole trapping stages, in a medical setting it can be the number of adverse events occurring during a follow up period or the number of hospitalizations. In this section we propose an alternative way of modeling capture-recapture count data via mixtures of Poisson when the mixing distribution is not constrained to belong to a parametric family. It is important to highlight the fact that, in this context, only the counts greater than zero can be observed.

In capture-recapture analyses, the count c_i represents the number of captures occurred to the same unit of the population during the whole experiment. However, there are other contexts where only positive counts are recorded and there is interest in inferring on the unobservable (missing) null counts. In estimating the number of species counts represent the number of units observed for each species (Chao & Lee 1992, Bunge & Fitzpatrick 1993) while in genetic studies they are associated to the number of times that a gene (Morris et al. 2003, Wang & Lindsay 2005, Thygesen & Zwinderman 2006) or clonetype (Sepulveda et al. 2010) has been reported as present or active. In both cases one is interested in the unreported null counts of unobserved species/genes/clonetypes. Hence, although we will mainly focus in

capture-recapture analysis, the procedures that we propose in this section can be implemented in other contexts which are concerned with positive count data with structurally unobservable null counts.

4.1 Poisson count data with individual heterogeneity

As in the previous chapter we conveniently label from 1 to M the actually observed units such that $c_i > 0$ and from $M+1$ to N the unobservable units such that $c_i = 0$. As discussed in the introduction, data representation binds in some way which models have to be considered. Indeed, when capture-recapture data are expressed as a series of counts and not in terms of individual binary capture history it is not possible to handle time varying behaviour of the catch rate . Hence models belonging to the classes $\mathcal{M}_T$ and $\mathcal{M}_B$ cannot be considered. One of the main feature of interest which one is left to model as flexibly as possible is the individual heterogeneity of the propensity of being captured. In fact, by including the possibility of having different parameters regulating the distribution of the individual count we are able to embed this important source of heterogeneity in our statistical model. In the following we will consider count data where the the maximum number of captures occurred is not a priori bounded. In capture-recapture context this kind of data can be thought of coming from a continuous-time experiment performed, say, during the interval $(0, t]$. Each count c_i can be considered as a realization of the final count of a continuous-time counting process C_i in $(0, t]$ with parameter λ_i representing the individual catch rate which is assumed to be constant during the whole trapping period. Given the central role of the Poisson distribution for count data we consider that

$$C_i \sim Poiss(\lambda_i) \qquad \forall i = 1, \ldots, N$$

We assume again that all units act independently from each others. The likelihood function could be expressed as follows

$$L(N, \lambda; \mathbf{c}) = \prod_{i=1}^{N} \frac{e^{-\lambda_i} \lambda_i^{c_i}}{c_i!}$$

where $\lambda = (\lambda_1, \ldots, \lambda_N)$, $\mathbf{c} = (c_1, \ldots, c_N)$. As pointed out above, in our capture-recapture context $c_{M+1}, \ldots, c_N$ correspond to unobservable null counts whose number $N - M$ can be determined once N is fixed and M units have been observed with positive count. As argued in Otis et al. (1978) the model described above conceptually involves $N + 1$ parameters: the parameter of interest N and N nuisance

parameters $\lambda_1, \ldots, \lambda_N$ which make the inferential problem partially indeterminate due to overparameterization. However, the rate parameter of the Poisson distribution for each unit can be thought of as an unobserved latent intensity and can be assumed to be drawn from a common distribution Q. Moreover, as discussed in the introduction, under the hypothesis of individual heterogeneity and hence for all models belonging to the class $\mathcal{M}_H$ the information conveyed by the counts c_i, $i = 1, \ldots, N$ can be summarized by the sufficient statistics called frequency of frequencies

$$\{n_k \, ; j = 1, \ldots T\}$$

where $n_k = \sum_{i=1}^{N} I(c_i = k)$ represents the number of units whose count corresponds to k and $T = \max(c_i)$. In a capture-recapture setting n_k represents the number of units with the same number k of captures during the whole experiment. Notice that as argued above the number n_0 of units with count equal to zero is not available but it is in one-to-one relation with N and $n_1, \ldots, n_T$, in fact

$$n_0 = N - \sum_{k=1}^{T} n_k = N - M$$

In a closed capture-recapture experiment as well as in other contexts, such as species richness problems, usually the main parameter of interest is the population size N and hence one needs to estimate the number n_0 i.e. to estimate the number of unobserved units.

In this hierarchical formulation the likelihood function can be written in terms of the sufficient statistics $\{n_k, k = 1, \ldots, T\}$ as follows

$$L(N, Q; \mathbf{n}) \propto \binom{N}{M} \prod_{k=0}^{T} [f(k, Q)]^{n_k} \tag{4.1}$$

where $\mathbf{n} = (n_1, \ldots, n_T)$ and Q is the mixing distribution for λ such that

$$f(k; Q) = \int_0^{\infty} \frac{e^{-\lambda} \lambda^k}{k!} dQ(\lambda). \tag{4.2}$$

In the literature alternative mixtures of Poisson with different finite (Pledger et al. 2003) or continuous (Böhning et al. 2005*a*) mixing distribution have been considered. Moreover a similar statistical setup has been recently considered in the species richness literature (Wang & Lindsay 2005, Mao & Lindsay 2007) and sometimes with completely unspecified count data distribution (Chao & Bunge 2002). In 2010, Wang proposed to consider a Poisson compound gamma model estimating the mixture by a nonparametric penalized maximum likelihood approach using a least-squares cross-validation procedure for the choice of the common shape parameter. Actually N is

considered the parameter of interest while Q is thought of as a nuisance parameter. If Q is not constrained to belong to a specific parametric family (4.1) is a nonparametric model for which several classical approaches have been already undertaken ranging from likelihood methods, jackknife or alternative coverage based estimator (see the recent overview in Wang (2010)) and none of them can be considered as a completely satisfactory solution. In this section we propose an alternative Bayesian approach which exploits the information resulting from the observed data integrating in a suitable way the likelihood in (4.1) using an appropriate prior distribution on N and a prior distribution on some essential features of Q. Our proposal yields an alternative nonparametric estimate of the population size based on integrated likelihood reparameterized in terms of a finite number of moments of a suitable mixing distribution.

4.2 Flexible moment modeling for unobserved individual heterogeneity

To begin with we show that, in order simplify our task, (4.1) can be approximated arbitrarily well by a model in which the mixing distribution Q has a compact support in $[0, u_\eta]$. In fact the following holds:
Theorem: Let Q be a generic probability distribution with support on $[0, \infty)$; $\forall\, \eta > 0\, \exists\, u_\eta > 0$ such that

$$d_{TV}\left(f(\cdot\,,\, Q), f(\cdot\,,\, Q_{u_\eta})\right) \leq \eta$$

where Q_{u_η} is the distribution Q restricted to have compact support on $[0, u]$
proof: In order to prove the theorem we have to verify that

$$\forall\, \eta > 0\, \exists\, u_\eta \,:\, |Q(A) - Q_{u_\eta}(A)| \leq \eta \quad \forall A \in \mathcal{B}(\mathcal{R}^+)$$

where $\mathcal{B}$ is the Borel σ-algebra. Since

$$Q_{u_\eta}(A) = \frac{Q\left(A \cap [0, u_\eta]\right)}{Q\left([0, u_\eta]\right)} \geq Q\left(A \cap [0, u_\eta]\right) \tag{4.3}$$

and

$$\forall\, \varepsilon(\eta) = \frac{\eta}{1+\eta} > 0\,;\, \exists\, u_\eta : Q([0, u_\eta]) > 1 - \varepsilon(\eta) \;\Rightarrow\; Q([0, u_\eta]^c) < \varepsilon(\eta) \tag{4.4}$$

we have

$$\begin{aligned} &Q(A) - Q_{u_\eta}(A) = Q\left(A \cap [0, u_\eta]\right) + Q\left(A \cap [0, u_\eta]^c\right) - Q_{u_\eta}(A) \leq \\ &Q\left(A \cap [0, u_\eta]\right) + Q\left([0, u_\eta]^c\right) - Q\left(A \cap [0, u_\eta]\right) < \varepsilon(\eta) < \eta \end{aligned}$$

Moreover, from (4.3) and (4.4) it follows that

$$Q_{u_\eta}(A) - Q(A) = \frac{Q\left(A \cap [0, u_\eta]\right)}{Q\left([0, u_\eta]\right)} - Q\left(A \cap [0, u_\eta]\right) + Q\left(A \cap [0, u_\eta]^c\right) \leq$$
$$\frac{Q\left(A \cap [0, u_\eta]\right)}{1 - \varepsilon(\eta)} - Q\left(A \cap [0, u_\eta]\right) \leq Q\left(A \cap [0, u_\eta]\right) \frac{\varepsilon(\eta)}{1 - \varepsilon(\eta)} \leq \eta$$

This minimal restriction on a compact support of the mixing distribution Q allows us to consider the one-to-one correspondence of a compact supported univariate distribution Q_u and the sequence of its moments. In fact we can simplify the functional form of the likelihood as a function of a finite number of characteristics of Q_u. To make it explicit we will be using first another one-to-one mapping between finite measures

$$dQ_u(\lambda) = e^\lambda dG_u(\lambda)$$

so that we can eventually regard the likelihood as a function of a finite number of moments of the finite measure $G_u(\cdot)$ uniquely corresponding to $Q_u(\cdot)$. In fact, for a fixed value u we can always consider the following simplified parametric model for the probability of each frequency counts

$$f(k; Q_u) = \int_0^u \frac{e^{-\lambda}\lambda^k}{k!} dQ_u(\lambda) = \frac{1}{k!}\int_0^u \lambda^k dG_u(\lambda) = \frac{m_k(G_u)}{k!} = f(k; G_u) \qquad (4.5)$$

where $m_k(G_u)$ is the k-th ordinary moment corresponding to the finite measure G_u not necessarily with total mass equal to 1. Indeed we can derive the corresponding likelihood

$$L(N, G_u; \mathbf{n}) \propto \binom{N}{M} \prod_{k=0}^{T} \left[f(k, Q_u)\right]^{n_k} = \binom{N}{M} \prod_{k=0}^{T} \left[\frac{m_k(G_u)}{k!}\right]^{n_k} \qquad (4.6)$$

which can be thought of as an approximate version of the original mixture of Poisson model (4.1). This suggests that the representation of the original model in terms of an infinite-dimensional functional parameter Q will be amenable to a flexible finite dimensional representation. This will ease the task of implementing a default Bayesian approach for making inference on the parameter of interest N.
Indeed, in order to further simplify the likelihood structure and make its expression to be a function of the moments of a *probability* measure (with fixed total mass equal to 1) supported on $[0, u]$ we will consider the following trick: we take the normalized probability distribution $\tilde{G}_u$ corresponding to G_u, namely

$$\tilde{G}_u(\cdot) = \frac{G_u(\cdot)}{\int_0^u dG_u(\lambda)}$$

so that

$$\begin{cases} m_0(\tilde{G}_u) = \int_0^u d\tilde{G}_u(\lambda) = 1 \\ \\ m_k(\tilde{G}_u) = \frac{m_k(G_u)}{m_0(G_u)} \end{cases}$$

It is immediate to realize that since $m_0(\tilde{G}_u) = 1$ we get

$$f(k, \tilde{G}_u) = \frac{1}{k!} \int_0^u \lambda^k d\tilde{G}_u(\lambda) = c \cdot f(k, Q_u) \quad \forall\, k = 0, \dots T$$

so that, summing up over all k the normalizing constant c is such that

$$c = \sum_{k=0}^{\infty} f(k, \tilde{G}_u) = \frac{1}{m_0(G_u)} = \frac{1}{f(0, G_u)} = \frac{1}{\int_0^u dG_u(\lambda)}.$$

One can replace the use of $f(k, Q_u)$ with $cf(k, \tilde{G}_u)$ and escape from the infinite summation defining from the latter expression a convenient approximation which represents a flexible parametric distribution for the frequencies of counts as follows

$$f(k, \boldsymbol{m}_{u,S}) = \frac{m_k(\tilde{G}_u)}{k! \sum_{j=0}^{S} \frac{m_j(\tilde{G}_u)}{j!}} \qquad k = 0, 1, \dots, S \tag{4.7}$$

where the probability parameters $f(k, \boldsymbol{m}_{u,S})$ are expressed as a function of the first S moments of the probability distribution $\tilde{G}_u$

$$\boldsymbol{m}_{u,S} = (m_{u,1}, \dots m_{u,k}, \dots, m_{u,S})$$

where

$$m_{u,k} = m_k(\tilde{G}_u) = \int_0^u \lambda^k d\tilde{G}_u(\lambda)$$

Usually $S = T$ but the parametric model is still well defined also for $S \neq T$. However, we point out that for the structure of the likelihood function (4.1) there is information only for the first T moments of the mixing distribution. The resulting model will be represented as

$$L(N, \boldsymbol{m}_{u,S}; \mathbf{n}) \propto \binom{N}{M} \prod_{k=0}^{S} \left[\frac{m_{u,k}}{k! \sum_{j=0}^{S} \frac{m_{u,j}}{j!}} \right]^{n_k} \tag{4.8}$$

and it can be considered a convenient approximation of (4.6) and hence of the original nonparametric model (4.1). We can make a final simplification by separating the dependence of $m_k(\tilde{G}_u)$ from u and the moments of a single probability distribution $\tilde{G}_1$ supported on $[0, 1]$ namely

$$m_k(\tilde{G}_u) = u^k m_k(\tilde{G}_1) \tag{4.9}$$

which corresponds to the change of measure for $\tilde{G}_u$ due to a scale factor u for the rate parameter λ. In the following we will use the notation m_k instead of $m_k(\tilde{G}_1)$ and $\mathbf{m}_S = (m_1, ..., m_S)$ will be the vector of the first S moments of an arbitrary probability distribution $\tilde{G}_1$ supported on $[0, 1]$. We can then express our flexible parametric model in terms of a vector of parameters $(N, \mathbf{m}_S, u) \in \{M, M+1, \dots\} \times \mathcal{M}_S \times [0, \infty)$ so that

$$L(N, \boldsymbol{m}_S, u; \mathbf{n}) \propto \binom{N}{M} \prod_{k=0}^{T} \left[\frac{u^k m_k}{k! \sum_{j=0}^{S} \frac{u^j m_j}{j!}} \right]^{n_k} \tag{4.10}$$

where the S- truncated moment space $\mathcal{M}_S$ is such that

$$\mathcal{M}_S = \left\{ (m_1, \dots, m_S) : m_k = \int_0^1 x^k d\tilde{G}_1(x) \ , \ \tilde{G}_1 \in \mathcal{P}([0, 1]) \right\}$$

where $\mathcal{P}([0, 1])$ is the class of probability distributions with support in $[0, 1]$. The ordinary moment space $\mathcal{M}_S$ is a constrained S-dimensional convex body and hence it is not easy to deal with. As proposed in Tardella (2002) and also used in Tardella & Farcomeni (2008) in the context of the discrete-time capture-recapture experiments one can also consider a further reparameterization of $\boldsymbol{m}_S$ in terms of the so-called canonical moments $\boldsymbol{c}_S = (c_1, \dots, c_S) \in [0, 1]^S$ (Skibinsky 1986, Dette & Studden 1997). We define the k-truncated moment class of distributions

$$\mathcal{P}_{\boldsymbol{m}_k} = \left\{ \tilde{G}_1 \in \mathcal{P}([0, 1]) : \int_0^1 x^r d\tilde{G}_1(x) = m_r \ , \ r = 1, \dots, k \right\}$$

where $\boldsymbol{m}_k = (m_1, \dots, m_k)$. Moreover, we define the following quantities

$$m_{k+1}^{+}(\boldsymbol{m}_k) = \sup_{\tilde{G}_1 \in \mathcal{P}_{\boldsymbol{m}_k}} m_{r+1}$$

$$m_{k+1}^{-}(\boldsymbol{m}_k) = \inf_{\tilde{G}_1 \in \mathcal{P}_{\boldsymbol{m}_k}} m_{r+1}$$

The generic element c_k of $\boldsymbol{c}_S$ is defined as follows

$$c_k = \frac{m_k - m_{k+1}^{-}(\boldsymbol{m}_k)}{m_{k+1}^{+}(\boldsymbol{m}_k) - m_{k+1}^{-}(\boldsymbol{m}_k)} \qquad k = 1, \dots, S$$

Then one can do all the computations and simulations in this unconstrained parameter space and finally reparameterize back into the space of the ordinary moments with little extra effort. so that MCMC approximations of the posterior distribution can be safely derived. In order to implement a fully Bayesian approach we need to set up a prior distribution for the vector of parameters involved in the model. In the next section we will give details on how to elicit a suitable prior distribution on the moment space $\mathcal{M}_S$.

4.3 Reference Bayesian inference

In order to implement a fully Bayesian approach for (4.10) we need to elicit the joint prior distribution for the whole parameter vector $(N, u, m_1, \ldots, m_S)$. We first show how a reference Bayesian inference can be derived for model (4.10) based on count frequency probabilities $f(k, \boldsymbol{m}_{u,S})$.
We note that, for fixed values of the parameters N and u taking $n_0 = N - \sum_{k=1}^{S} n_k$ the expression in (4.10) is a multinomial likelihood in terms of the probabilities $f(0, \boldsymbol{m}_{u,S}), \ldots, f(S, \boldsymbol{m}_{u,S})$ which are in turn one-to-one related to $m_0, \ldots, m_S$. This allows us to consider a standard Jeffreys'prior on $f(0, \boldsymbol{m}_{u,S}), \ldots, f(S, \boldsymbol{m}_{u,S})$ and transform it back in terms of a default distribution on $m_0, \ldots, m_S$ for any fixed value of N and u taking into account the appropriate Jacobian. It is known that the Jeffreys'prior for an unconstrained multinomial parameter vector is a Dirichlet distribution and one can argue that for the count frequency probabilities which are constrained on a proper convex body contained in the S-dimensional simplex the same functional form of the Jeffreys' prior is preserved up to a different normalizing constant. So we have

$$\pi_J(f(1; \boldsymbol{m}_{u,S}), \ldots, f(1; \boldsymbol{m}_{u,S})) \propto \prod_{k=0}^{S} [f(k; \boldsymbol{m}_{u,S})]^{-\frac{1}{2}} \tag{4.11}$$

As previously mentioned simulation within the moment space can be eased reparameterizing the ordinary moments of the distribution $\tilde{G}_1 \in [0, 1]$ in terms of the corresponding canonical moments (Tardella 2002). The only step needed to re-express our Jeffreys prior in terms of $m_1, \ldots, m_S$ is the evaluation of the appropriate Jacobian. Indeed, to simplify formulae, let us denote with $x_k = f(k, \boldsymbol{m}_{u,S})$, $y_k = \frac{m_k(\tilde{G_u})}{k!}$, $\mathbf{x} = (x_1, \ldots, x_S)$ and $\mathbf{y} = (y_1, \ldots, y_S)$. The count frequencies in (4.7) can be expressed as a function of $\mathbf{y}$:

$$\mathbf{x} = g(\mathbf{y})$$

as follows

$$x_k = \frac{y_k}{\sum_{j=0}^{S} y_j} = \frac{y_k}{D_{\mathbf{y}}}$$

where $D_{\mathbf{y}} = \sum_{j=0}^{S} y_j$ stands for the denominator. Notice that both vectors $\mathbf{x}$ and $\mathbf{y}$ can be completed when needed by $x_0 = f(0, \boldsymbol{m}_{u,S})$ and $y_0 = \frac{m_0(\tilde{G_u})}{0!}$ using the known constraints: $\sum_{k=0}^{S} x_k = 1$ and $y_0 = 1$. Hence we have that the standard Jeffreys'prior on multinomial cell probabilities $\mathbf{x}$ is

$$\pi_J(\mathbf{x}) \propto \prod_{k=0}^{S} x_k^{-\frac{1}{2}}$$

and the corresponding Jeffreys'prior in terms of $\mathbf{y} = g^{-1}(\mathbf{x})$ can be written as

$$\pi_J^\star(\mathbf{y}) = \pi_J(g(\mathbf{y}))\cdot \mid J_g(\mathbf{y}) \mid \tag{4.12}$$

where $\mathbf{J}_g(\mathbf{y}) = [j_{i,j}(\mathbf{y})]$ is the Jacobian matrix containing the partial derivatives of $g(\mathbf{y})$. The Jacobian matrix has the extra-diagonal elements

$$j_{i,j}(\mathbf{y}) = -\frac{y_j}{D_{\mathbf{y}}^2} \qquad \forall i\, \forall j;\ i \neq j$$

while the diagonal elements are

$$j_{i,i}(\mathbf{y}) = \frac{D_{\mathbf{y}} - y_i}{D_{\mathbf{y}}^2} \qquad i = 1, \ldots, S$$

Now we finally express the Jeffreys'prior in terms of $\mathbf{m}_S$ using (4.12) and the one-to-one mapping (4.9) which maps $\mathbf{y}$ into $\mathbf{m}_S$

$$y_k = \frac{u^k}{k!} m_k \Rightarrow \mathbf{y} = h(mathbf m_S)$$

and hence we have

$$\pi_R(\mathbf{m}_S) = \pi_J(g(h(\mathbf{m}_S))) \cdot \mathbf{J}_g(h(\mathbf{m}_S))\cdot \mid \mathbf{J}_h(\mathbf{m}_S) \mid$$

where $\mid \mathbf{J}_h(\mathbf{m}_S) \mid$ is easily to derived as follows

$$\mid \mathbf{J}_h(\mathbf{m}_S) \mid = \prod_{k=1}^{S} \frac{u^k}{k!}$$

To complete the prior elicitation for our model we consider for N, similarly to what has been done in Part 2, three different non-informative prior distributions: uniform, $1/N$ and Rissanen's prior. We will investigate the sensitivity of the posterior analyses and compare its performances by simulation study and results of some real data examples.

Notice that so far we have assumed a fixed upperbound u for the support of the mixing distribution of λ. Now we need to endow u with a prior distribution. Indeed considering how we jointly rescale all the moments of $\tilde{G}_1$ into the moments of $\tilde{G}_u$

$$\begin{aligned} & m_1(\tilde{G}_u) = u\, m_1(\tilde{G}_1) \\ & \ldots \\ & m_k(\tilde{G}_u) = u^k\, m_1(\tilde{G}_1) \\ & \ldots \\ & m_S(\tilde{G}_u) = u^S\, m_1(\tilde{G}_1) \end{aligned}$$

we use as a reference distribution

$$\pi_R(u) \propto u^{-\frac{s(s+1)}{2}} \tag{4.13}$$

In order to avoid an improper distribution and degenerate inference for $u \to 0$ we fix a positive lowerbound ($u_{LB} = 0.5$) for the support of u.

4.4 Simulated data

In order to evaluate the performance of our proposal we implemented a simulation study according to the same setting considered in Wang (2010) as described in Table 4.1. For each setting a different mixing distribution on the Poisson intensity is fixed

Setting	**Distribution** (Q)	$E(M/N)$
	Gamma	
1	$Ga(4, 3.125)$	0.90
2	$Ga(4, 1)$	0.59
3	$Ga(1, 0.25)$	0.20
	Gamma Mixture	
4	$0.5 \cdot Ga(2, 1) + 0.5 \cdot Ga(2, 2)$	0.65
5	$0.5 \cdot Ga(2, 1) + 0.5 \cdot Ga(4, 1)$	0.57
	Log-Normal	
6	$LN(0.75, 0.75)$	0.82
7	$LN(-0.5, 2)$	0.50
8	$LN(-1, 1)$	0.36
	Log-Normal Mixture	
9	$0.5 \cdot LN(-0.5, 1) + 0.5 \cdot LN(0.5, 1)$	0.61
	Finite Mixture	
10	$0.8 \cdot\delta(1.2) + 0.2 \cdot \delta(6.7)$	0.76
11	$0.89 \cdot\delta(0.5) + 0.11 \cdot \delta(6.7)$	0.46
12	$0.8 \cdot\delta(0.2) + 0.2 \cdot \delta(1.3)$	0.29

Table 4.1: *Simulation setting (Wang (2010))*

and 100 simulated datasets are drawn and used to repeat the estimation procedure. Bias and mean square error of point estimates and coverage of interval estimates are approximatively evaluated averaging the results obtained with the simulated datasets. We compare our method with the recent non parametric approach based on a penalized likelihood proposed in Wang (2010) which highlighted inferential difficulties of the previously available approaches and showed a substantial improvement over the latter. Wang's procedure is implemented in the `R` package `SPECIES` (Wang 2011) where the corresponding function is named `pcg(...)`. The package allows also to compute point and confidence interval estimates from alternative nonparametric and semi-parametric methods using the first S counts observed. In order to

make a sound comparison with Wang's procedure we fixed the number of moments of the probability distribution $\tilde{G}_u$ considered to be $S = 10$ since in Wang's simulation study only the first 10 counts are considered. Although we evaluated several prior choices for N we report in Table 4.2 only the results obtained from the uniform prior $\pi(N) \propto 1$ which leads to the best performances. 5On the other hand we will show the results from both prior choices for u: $\pi_R(u)$ and $\pi_A(u)$. We will denote by $\hat{N}_{BPM}$ the resulting estimator. As we can see from the results in Table 4.2

Setting	$\hat{N}$	$\hat{Me}$	MSE	$\%\,Cov$	**Setting**	$\hat{N}$	$\hat{Me}$	MSE	$\%\,Cov$
1	$\hat{N}_{BPM}$	1020	27.93	100	**2**	$\hat{N}_{BPM}$	1135	160.73	99
	$\hat{N}_{PL}$	1020	28.11	97		$\hat{N}_{PL}$	1138	161.00	99
	$\hat{N}_{PCG}$	1011	28.39	95		$\hat{N}_{PCG}$	1014	149.47	99
3	$\hat{N}_{BPM}$	1070	147.85	100	**4**	$\hat{N}_{BPM}$	1009	58.08	100
	$\hat{N}_{PL}$	1034	133.25	100		$\hat{N}_{PL}$	1013	59.16	100
	$\hat{N}_{PCG}$	924	234.71	100		$\hat{N}_{PCG}$	991	124.47	99
5	$\hat{N}_{BPM}$	1041	72.02	100	**6**	$\hat{N}_{BPM}$	1004	106.64	100
	$\hat{N}_{PL}$	1040	72.42	100		$\hat{N}_{PL}$	997	102.60	100
	$\hat{N}_{PCG}$	1009	160.21	96		$\hat{N}_{PCG}$	1041	113.63	98
7	$\hat{N}_{BPM}$	829	171.03	83	**8**	$\hat{N}_{BPM}$	907	113.89	100
	$\hat{N}_{PL}$	831	169.51	86		$\hat{N}_{PL}$	912	115.77	100
	$\hat{N}_{PCG}$	996	198.86	97		$\hat{N}_{PCG}$	1016	197.61	99
9	$\hat{N}_{BPM}$	976	71.94	98	**10**	$\hat{N}_{BPM}$	1117	122.48	72
	$\hat{N}_{PL}$	974	71.88	97		$\hat{N}_{PL}$	1061	78.02	88
	$\hat{N}_{PCG}$	1028	163.07	100		$\hat{N}_{PCG}$	1038	56.93	83
11	$\hat{N}_{BPM}$	1207	281.11	91	**12**	$\hat{N}_{BPM}$	880	154.43	100
	$\hat{N}_{PL}$	1192	276.01	91		$\hat{N}_{PL}$	879	153.87	100
	$\hat{N}_{PCG}$	1035	177.26	87		$\hat{N}_{PCG}$	938	169.39	93

Table 4.2: *Comparing four different estimators with respect to median bias, mean squared error and 95% confidence interval coverage in 12 simulation settings listed in Table 4.1*

graphically summarized in Figure 4.1 our Bayesian estimators seem to compete well with Wang's `pcg` procedure although occasionally they can be beaten in terms of efficiency and interval coverage. In his paper Wang shows how his estimator almost uniformly outperforms all previously available estimators in terms of precision and coverage. We find out that a slight modification of the fully Bayesian recipe can do

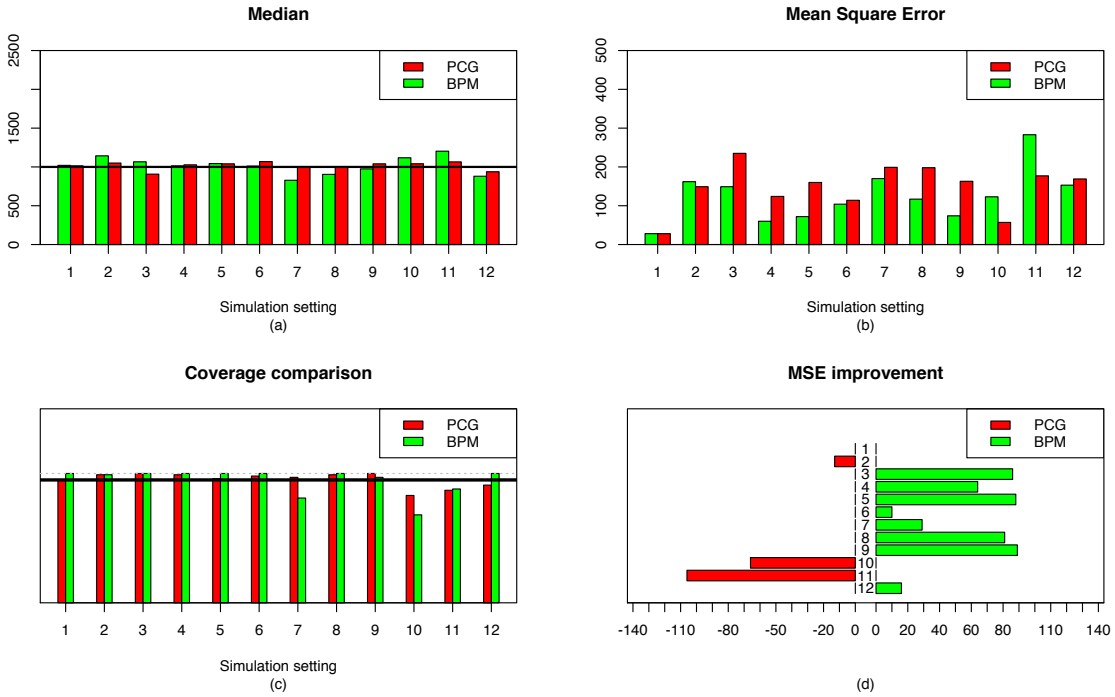

Figure 4.1: *Comparing PCG and fully Bayesian approach: Summary*

even better. It turns out that integrating out the following penalized likelihood

$$L_P(N, \boldsymbol{m}_S, u; \mathbf{n}) \propto \binom{N}{M} \prod_{k=0}^{T} \left[\frac{u^k m_k}{k! \sum_{j=0}^{S} \frac{u^j m_j}{j!}} \right]^{n_k - \frac{1}{2}}$$

with the similar prior choices for N and u and a uniform measure on the moments $m_1, \ldots, m_S$ one gets a better performance as we can see in Figure 4.2. However, we will not consider it further because it does not correspond to a fully Bayesian approach.

Moreover, even though our new methods (fully Bayesian and penalized integrated likelihood) are computationally intensive, the derivation of the interval estimates is often quicker compared to Wang's `pcg` procedure which relies on a costly double-bootstrap procedure. Overall if we average on all the twelve simulation settings our N_{BPM} turns out to be an improvement over N_{PCG} in terms of average mean square error while the corresponding interval estimates show an overall suitable coverage close to the nominal level.

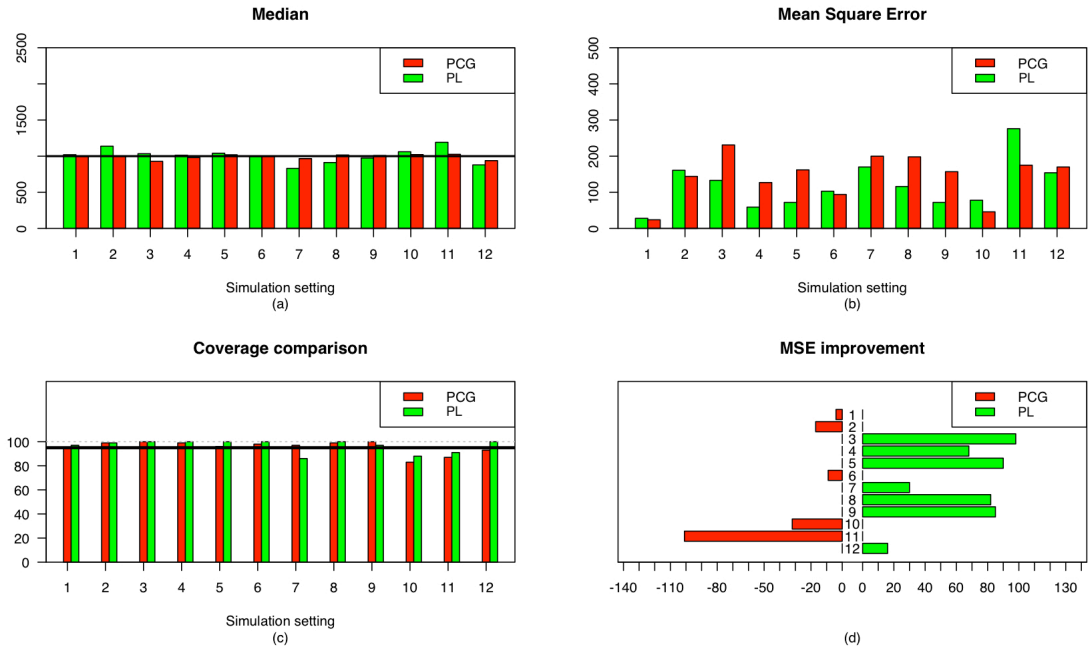

Figure 4.2: *Comparing PCG and integrating a modified/penalized likelihood approach: Summary*

4.5 Real data analyses

We investigate the effectiveness of our proposed estimator with several benchmark datasets used in the recent works of Wang (2010) and Rocchetti et al. (2011) comparing our Bayesian approach with both approaches developed in these papers. The estimator $\hat{N}_{RBB}$ proposed in Rocchetti et al. (2011) is based on a linear regression model on the ratios of successive frequency counts. Namely

$$\hat{r}(x) = \frac{(x+1)n_{x+1}}{xn_x}$$

We stress that such estimator does not aim to be a flexible nonparametric estimator since it is derived under the assumption that the count distribution belongs to the so called Katz family (Katz 1952). For this reason we have not used it as alternative competitor in our simulation study. For the following real data Bayesian analyses we will follow the recipe recommended from the simulation study: uniform prior for N, Jeffreys'prior on $\boldsymbol{m}_S$ and for u we consider the reference prior $\pi_R(u)$ described in (4.13).

Traffic data

We start with the famous dataset known as *Traffic Data* originally studied in Simar (1976) and lately re-analyzed in Böhning et al. (2005*b*) and Wang (2010). Data are shown in Table 4.3. They represent the accident counts submitted to La Royale Belge Insurance Company during a particular year. In this example we know the real value for N (9461) which is the total number of insurance policies covering both "business" and "tourist" automobiles; hence the complete frequency counts show that the proportion of the unobserved units is very high. For the analysis we

k	**1**	**2**	**3**	**4**	**5**	**6**	**7**	M
Traffic (n_k)	1317	239	42	14	4	4	1	1621

Table 4.3: *Traffic data-frequencies*

have considered all the available positive counts $n_1, \dots n_S$ with S equal to 7 which is indeed the maximum count observed. The MCMC algorithm runs for 110000 iterations discarding the first 10000. In Figure 4.3 the trace plots of the three main quantities: N, u and m_1 are shown. It is apparent that there is a strong autocorrelation which is likely yielding a slow mixing of the chain and can affect the resulting Monte Carlo error.

This strong autocorrelation can be due to the strong dependence among the three main quantities as evidenced from the scatter plots in Figure 4.4 (especially the one corresponding to N and m_1). However, we have verified that the results do not vary appreciably with a larger MCMC size. Indeed we redraw the acf considering a thin factor $\psi = 50$ leading 2000 iterations. The resulting acf in Figure 4.5 looks reasonable. As far as inference on N is concerned we can see from the histogram in Figure 4.6 that the known value $N = 9461$ is also very close to the mode of the posterior distribution of N. In Table 4.4 are expressed point and interval estimates from different prior choices of N and u. As we can see the point estimates are sufficiently stable with respect to the prior choice strategy. Moreover our credible intervals always contain the true N although the sensitivity of the upper bound of the credible intervals seems to be more pronounced than in the case of point estimates.
When we compute alternative estimators $\hat{N}_{PCG}$ proposed in Wang (2010) and $\hat{N}_{RBB}$ proposed in Rocchetti et al. (2011) we have that both seem to be more conservative and underestimate somehow the true N (6935 and 7840 respectively). However, in Wang (2010) among many alternative classical procedures considered in that paper only the confidence interval derived from $\hat{N}_{PCG}$ through a double-bootstrap

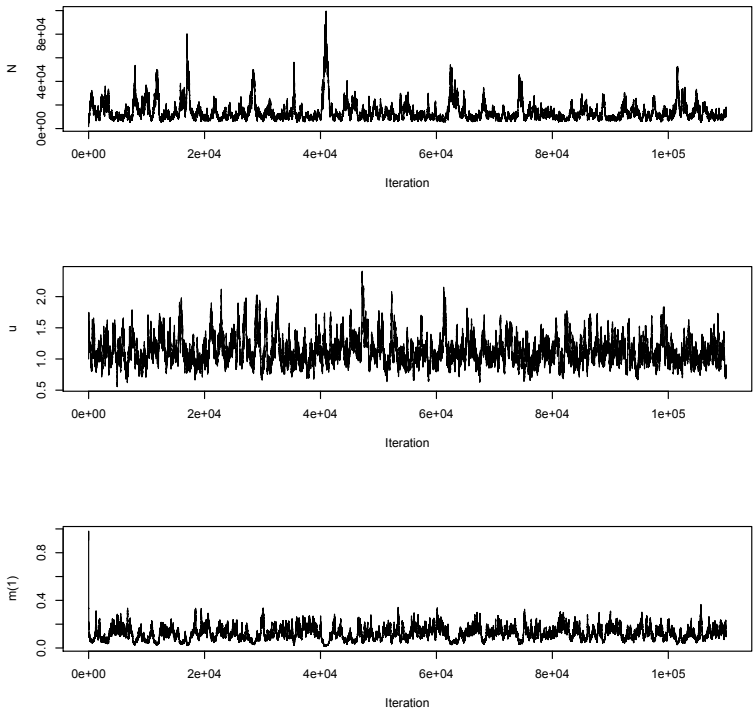

Figure 4.3: *Trace-plot of N, u and $m_{1,7}$.*

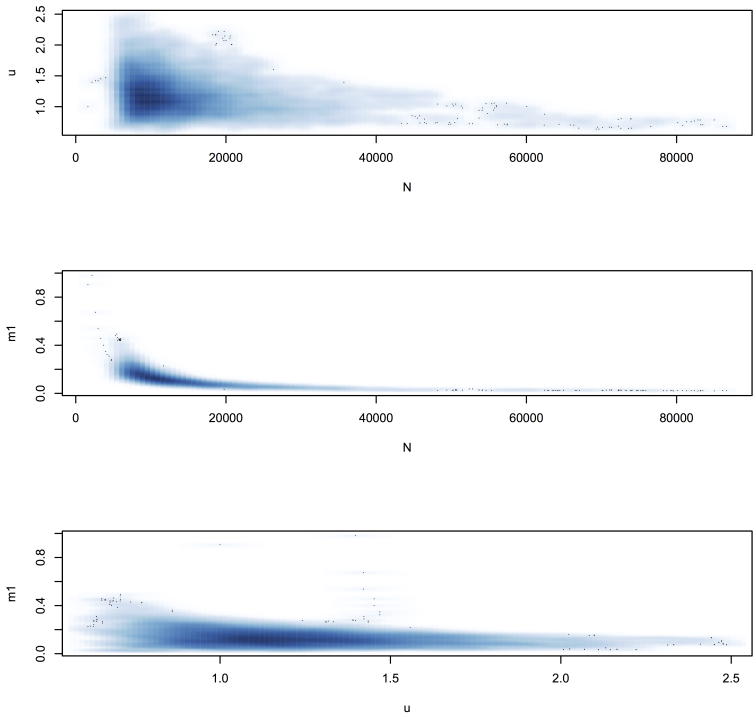

Figure 4.4: *Scatter plot of N, u and $m_{1,7}$.*

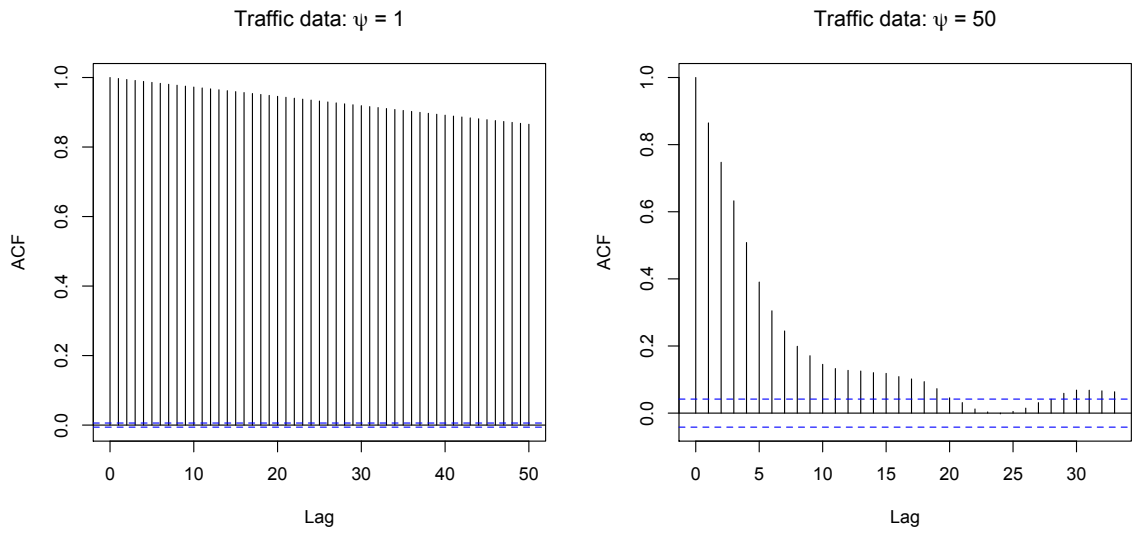

Figure 4.5: *Traffic data: acf of N with thin factor $\psi = 1, 50$.*

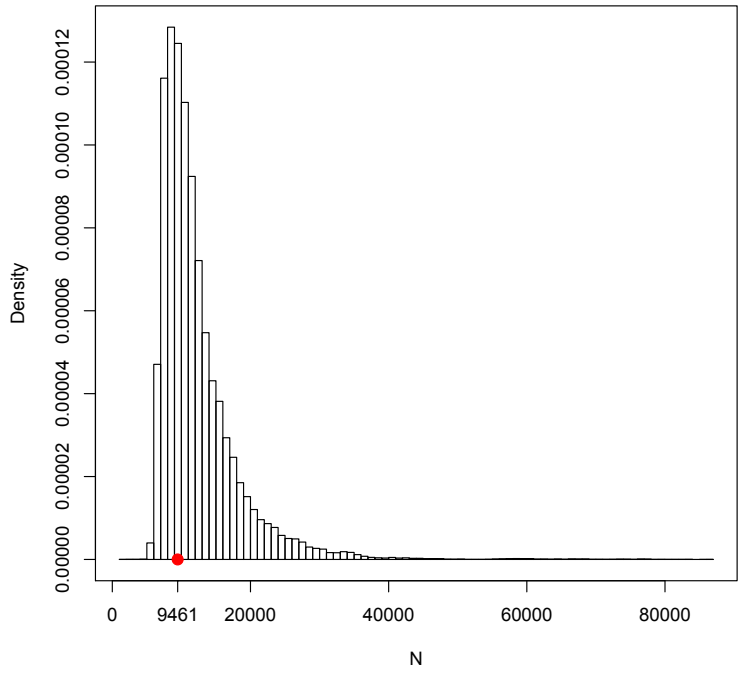

Figure 4.6: *Traffic data: Histogram of MCMC samples from the posterior distribution of N.*

procedure gets the true N inside. Hence we consider our estimator of N in this example one of the few successful estimators of the quantity of interest, in fact the

Methods	$\hat{N}$	N^-	N^+
BPM	9548	5642	22582
$BPM_{\frac{1}{N}}$	9121	5416	22816
$BPM_{Rissanen}$	8970	5662	18255
PCG	6935	5121	12843
RBB	7840	7742	7937

Table 4.4: *Traffic data: alternative point and interval estimates*

closest one to the true known value.

Root data

In Table 4.5 are shown the *Root data* already analyzed in Wang (2010) which represent the count distribution of the expressed genes of the arabidopsis thaliana in the root tissue. Notice that in this case there is a genuine interest in the unknown number of unexpressed genes since data are collected from a cDNA library sample which, very likely does not allow a full screening of all expressed genes.

k	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	
Root (n_k)	2187	490	133	121	37	51	22	19	7	
	10	**11**	**12**	**13**	**14**	**15**	**16**	**17+**		M
	8	6	7	6	4	5	5	18		3126

Table 4.5: *Root data-frequencies*

Researchers agreed that the arabidopsis thaliana has a relatively small genome with approximatively 27000 protein coding genes not necessarily all expressed in all tissues. This information can be easily exploited in our Bayesian procedure formalizing an ad-hoc prior distribution for N by setting a suitable upperbound for the population size of the expressed genes. We fix $N_{upp} = 30000$ for our analysis. On the other hand this (a priori) information cannot be employed so easily in the alternative classical approaches.
The results of the three alternative procedures are shown in Table 4.6. As we can see the point estimates $\hat{N}_{PCG}$ and $\hat{N}_{RBB}$ are very close together (8980 and 8870 respectively). As argued in Wang (2010) they could be a conservative estimate of the total number of expressed genes in the root tissue. Our estimate is considerably higher exceeding the value 11000 for both prior choices. Although in this case

Methods	$\hat{N}$	N^-	N^+
BPM	11073	8739	15316
PCG	8980	8383	18771
RBB	8970	8652	9288

Table 4.6: *Root data: alternative point and interval estimates*

the population size is not known in advance, however previous works (Ma et al. 2005) suggest a percentage of expressed genes in root tissue greater than 40% of the 27000 protein coding genes and which fits well with the recommendation provided by $\hat{N}_{BPM}$.

Colorectal polyps

From medical research experiences it is well recognized that diagnosing adenomatous polyps can be subjected to undercount due to misclassification at colonoscopy. We use data from Alberts et al. (2000) where in order to evaluate the recurrence of colorectal adenomatous polyps subjects with previous history of colorectal adenomatous polyps are allocated to one of two treatment groups, low fiber and high fiber. Polyps data-frequency distribution of recurrent adenomatous polyps per patient, by treatment group is reported in Table 4.7. For both groups the population size is known in advance: 584 for the low fiber treatment ($n_0 = 285$) and 722 for high fiber treatment ($n_0 = 381$) respectively.

k	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12+**	M
Polyps low (n_k)	145	66	39	17	8	8	7	3	1	0	2	3	299
Polyps high (n_k)	144	61	55	37	17	5	4	6	5	1	1	5	341

Table 4.7: *Polyps data-frequency distribution*

In Table 4.8 are reported alternative point and the interval estimates for both treatments. In this case Wang's estimator gets closer to the true N and also its confidence intervals include the main parameters of interest. Notice that, differently from the other procedures it overestimate the true population size.
Our proposal, although slightly negatively biased, yields confidence intervals which always contain the true N for both data sets and they are also narrower than those resulting from Wang's approach. Moreover, as we can see from the acf plots in Figure 4.7 the autocorrelation is sensibly lower with respect to the *Traffic data* example.

	Methods	$\hat{N}$	N^-	N^+
Polyps low	BPM	521	410	717
	PCG	626	424	780
	RBB	492	446	534
Polyps high	BPM	544	429	758
	PCG	806	526	956
	RBB	496	425	567

Table 4.8: *Polyps-data: alternative point and interval estimates*

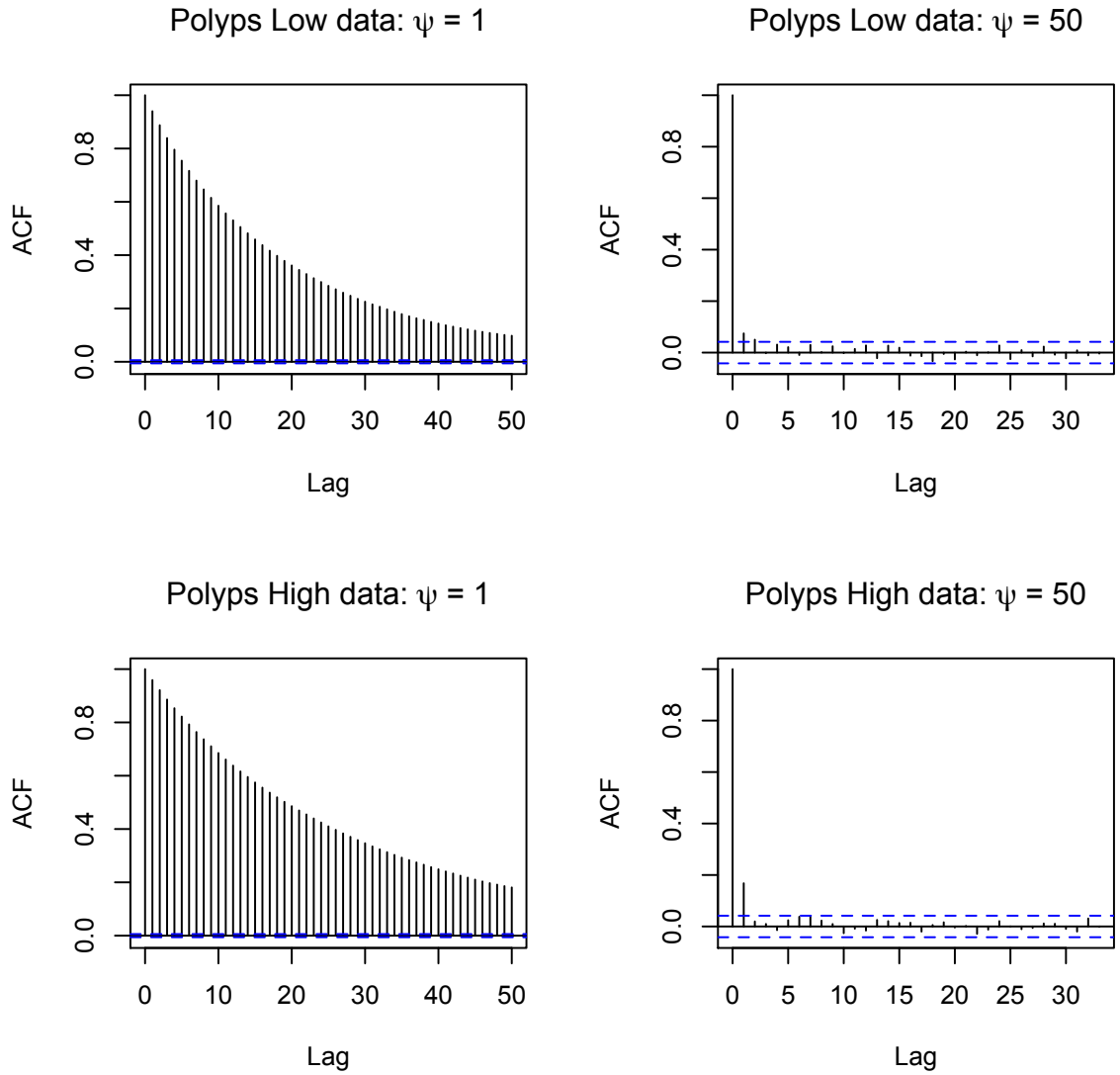

Figure 4.7: *Polyps low-high data: acf of N with thin factor $\psi = 1, 50$.*

Scrapie in Great Britain (2002-2006)

In Great Britain, scrapie is an endemic fatal neurological disease which affects small ruminants (e.g. sheep, goats etc). In Table 4.9 is reported the distribution of counts of confirmed scrapie-affected sheep in Great Britain between 2002 and 2006 Rocchetti et al. (2011). For all procedures we consider the truncated distribution of the

k	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10+**	M
Scrapie (n_k)	298	89	42	17	20	7	11	7	3	22	516

Table 4.9: *Scrapie data-frequencies*

the first 9 counts while the frequencies n_k corresponding to the counts $k \geq 10$ are summed up to the resulting estimates. As we can see from Table 4.10 the estimates produced by $\hat{N}_{BPM}$ and $\hat{N}_{RBB}$ are close together (1269 and 1220 respectively). However, our procedure yields wider confidence interval compared with RBB recognizing the possibility of more than 1500 cases of scrapie. On the other hand, the estimates obtained by the Poisson-compound gamma approach of Wang appear much higher than the alternative estimators ($\hat{N}_{PCG} = 1993$) and somehow surprisingly high with respect to other recent analyses with the same data set (Böhning et al. 2011). Indeed, the corresponding completeness rate of 25.9% seems to be too low in this case. Notice, however, that the point estimate returned by `pcg` is not incompatible with our Bayesian inference in terms of its credible interval. On the other hand, the interval estimate returned by `pcg` function in `SPECIES` package looks inconsistently beyond the point estimate possibly due to some numerical instability problems.

Methods	$\hat{N}$	N^-	N^+
BPM	1269	890	2165
PCG	1993	4312	13638
RBB	1220	1151	1289

Table 4.10: *Scrapie data: alternative point and interval estimates*

Methamphetamine use in Thailand

Data in Table 4.11 is concerned with the drug abuse in Thailand during the last quarter of 2001. In this table the number of methamphetamine users are displayed for each count of treatment episodes reported by the public health surveillance system. A total of 3345 distinct drug users have been observed with maximum

k	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	M
Methamphetamine (n_k)	3114	163	23	20	9	3	3	3	4	3	3345

Table 4.11: *Methamphetamine data-frequencies*

number of captures T equal to 10. The count distribution has a very strongly positive skewness: 3114 out of 3345 units present only one capture. This is a clue for a severe undercount or, which is the same, a large frequency n_0 of unreported users. The point estimates from Wang and B-B-R are 55739 and 61133 respectively.

Methods	$\hat{N}$	N^-	N^+
BPM	55435	35472	109171
PCG	55739	34783	93658
RBB	61133	60986	61280

Table 4.12: *Methamphetamine data: alternative point and interval estimates*

As reported in Table 4.12 our point estimate is only slightly lower ($N_{BPM} = 55435$). However, similarly to Wang's procedure, our confidence interval confirms that there can be more than 100000 drug users. Moreover, the lower limits of the of the interval is very close to Chao's lower bound

$$\hat{N}_{C.lb} = M + \frac{n_1^2}{2\, n_2} = 33090$$

which is a conservative nonparametric estimator based on the Cauchy-Schwarz inequality.

4.6 Final remarks

In this section we have dealt with modeling individual heterogeneity within Poisson count distribution in the absence of zero counts. We developed an original flexible approximation of a mixture of Poisson distributions where the mixing distribution is not constrained to belong to a specific parametric family.

Our Bayesian approach described in Section 4.2 and 4.3 is based on a reparameterization of the mixture likelihood function (4.1) in terms of the first S ordinary moment corresponding to a finite measure G_u with support $[0, u]$ where u is not necessarily fixed. In order to obtain a probability measure with total mass equal to 1 we have rescaled G_u to $\tilde{G}_u$ and then we have truncated the infinite sequence

of moments of $\tilde{G}_u$ to the first S moments using an explicit renormalization which formally resembles the original likelihood (4.6). Moreover, we have exploited the reparameterization of the ordinary moments into the so-called canonical moments conveniently rescaled in $[0, 1]$ allowing for an easier MCMC implementation. Finally, in order to set-up an appropriate prior distribution on the moment space we noted that conditionally on N and u the likelihood function has a multinomial structure which allows us to consider a standard Jeffreys'prior opportunely expressed in terms of moments with the appropriate Jacobian.

Formal arguments and a simulation study suggested a reference Bayesian recipe corresponding to a uniform prior for N and an invariant prior for u as described in (4.13). As shown from the simulation results our new fully Bayesian approach seems to perform well in terms of efficiency and coverage although slightly more biased than Wang's estimates. The good performances of the proposed Bayesian procedure are also confirmed from the results obtained in several real data analyses where our Bayesian approach always produced reasonable values for both point and interval estimates. Indeed for data sets where it is known in advance the population size (Traffic and Polyps data) the point estimates were close to the truth and the interval estimates always contained to the true value of N while for the other data-sets our proposal well agreed with previous scientific knowledge of the corresponding phenomenon.

The acf plots highlighted sometimes slow convergence. However results obtained by our Bayesian procedure seem to be sufficiently stable and reliable. Our analysis is computationally more intensive than Wang's procedure for point estimates but lighter for interval estimates since it relies on a costly bootstrap procedure.

As future work, it would be interesting to explore the asymptotic behaviour of the procedures for $N \to \infty$. As argued in Mao & Lindsay (2007), we do not have to expect good results from conditional likelihood approach, especially in terms of the coverage of the interval estimates. However, in the examples proposed for N in the range of thousands our estimates behave reasonably well and candidates itself to be a good alternative to the recent N_{PCG} estimator recently proposed by Wang.

Bibliography

Alberts, D. S., Martinez, M. E., Roe, D. J., Guillen-Rodriguez, J. M., Marshall, J. R., van Leeuwen, J. B., Reid, M. E., Ritenbaugh, C., Vargas, P. A., Bhattacharyya, A. B., Earnest, D. L. & Sampliner, R. E. (2000), 'Lack of effect of a high-fiber cereal supplement on the recurrence of colorectal adenomas. Phoenix Colon Cancer Prevention Physicians' Network', *N. Engl. J. Med.* **342**(16), 1156–1162.

Alho, J. M. (1990), 'Logistic regression in capture-recapture models', *Biometrics* **46**, 623–635.

Amstrup, S. C., McDonald, T. L. & Manly, B. F., eds (2005), *Handbook of Capture-Recapture Analysis*, Princeton University Press.

Bartolucci, F. & Pennoni, F. (2007), 'A Class of Latent Markov Models for Capture-Recapture Data Allowing for Time, Heterogeneity, and Behavior Effects', *Biometrics* **63**(2), 568–578.

Bishop, Y. M. M., Fienberg, S. E. & Holland, P. W. (1975), *Discrete Multivariate Analyses: Theory and Practice*, MIT Press.

Böhning, D., Dietz, E., Kuhnert, R. & Schön, D. (2005*a*), 'Mixture models for capture-recapture count data', *Stat. Methods Appl.* **14**(1), 29–43.

Böhning, D., Dietz, E., Kuhnert, R. & Schön, D. (2005*b*), 'Mixture models for capture-recapture count data', *Stat. Methods Appl.* **14**(1), 29–43.

Böhning, D., Kuhnert, R. & Vilas, V. D. R. (2011), 'Capture-recapture estimation by means of empirical Bayesian smoothing with an application to the geographical distribution of hidden scrapie in Great Britain', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **60**(5), 723–741.

Bunge, J. & Fitzpatrick, M. (1993), 'Estimating the number of species: A review', *Journal of the American Statistical Association* **88**, 364–373.

Carle, F. L. & Strub, M. R. (1978), 'A new method for estimating population size from removal data', *Biometrics* **34**, 621–630.

Chaiyapong, Y. & Lloyd, C. J. (1997), 'Accurate inference for recapture experiments with behavioural response', *Journal of Statistical Computation and Simulation* **56**, 97–115.

Chao, A. (2001), 'An overview of closed capture-recapture models', *Journal of Agricultural, Biological and Environmental Statistics* **6**(2), 158–175.

Chao, A. & Bunge, J. (2002), 'Estimating the number of species in a stochastic abundance model', *Biometrics* **58**(3), 531–539.

Chao, A., Chu, W. & Hsu, C.-H. (2000), 'Capture-recapture when time and behavioral response affect capture probabilities', *Biometrics* **56**(2), 427–433.

Chao, A. & Lee, S.-M. (1992), 'Estimating the number of classes via sample coverage', *J. Amer. Statist. Assoc.* **87**(417), 210–217.

Dette, H. & Studden, W. J. (1997), *The theory of canonical moments with applications in statistics, probability, and analysis*, John Wiley & Sons Inc., New York. A Wiley-Interscience Publication.

Farcomeni, A. (2011), 'Recapture models under equality constraints for the conditional capture probabilities', *Biometrika* .

Fattorini, L., Marcheselli, M., Monaco, A. & Pisani, C. (2007), 'A critical look at some widely used estimators in mark-resighting experiments', *Journal of Animal Ecology* **76**(5), 957–965.

Ghosh, S. K. & Norris, J. L. (2005), 'Bayesian capture-recapture analysis and model selection allowing for heterogeneity and behavioral effects', *Journal of Agricultural, Biological, and Environmental Statistics* **10**(1), 35–49.

Huggins, R. & Hwang, W.-H. (2011), 'A review of the use of conditional likelihood in capture-recapture experiments', *International Statistical Review* **79**(3), 385–400.

Huggins, R. M. (1989), 'On the statistical analysis of capture experiments', *Biometrika* **76**, 133–140.

Huggins, R. M. (1991), 'Some practical aspects of a conditional likelihood approach to capture experiments', *Biometrics* **47**, 725–732.

Hwang, W.-H., Chao, A. & Yip, P. S. F. (2002), 'Continuous-time capture-recapture models with time variation and behavioural response', *Australian & New Zealand Journal of Statistics* **44**(1), 41–54.

Hwang, W.-H. & Huggins, R. (2011), 'A semiparametric model for a functional behavioural response to capture in capture-recapture experiments', *Australian & New Zealand Journal of Statistics* **53**(4), 403–421.

Katz, L. (1952), 'The distribution of the number of isolates in a social group', *The Annals of Mathematical Statistics* **23**(2), pp. 271–276.

Lang, J. B. (1996), 'Maximum likelihood methods for a generalized class of log-linear models', *The Annals of Statistics* **24**(2), 726–752.

Lee, S.-M. & Chen, C. W. S. (1998), 'Bayesian inference of population size for behavioral response models', *Statistica Sinica* **8**, 1233–1248.

Lee, S.-M., Hwang, W.-H. & Huang, L.-H. (2003), 'Bayes estimation of population size from capture-recapture models with time variation and behavior response', *Statistica Sinica* **13**(2), 477–494.

Link, W. A. (2003), 'Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities', *Biometrics* **59**(4), 1123–1130.

Link, W. A., Yoshizaki, J., Bailey, L. L. & Pollock, K. H. (2010), 'Uncovering a latent multinomial: Analysis of mark-recapture data with misidentification', *Biometrics* **66**(1), 178–185.

Ma, L., Sun, N., Liu, X., Jiao, Y., Zhao, H. & Deng, X. W. (2005), 'Organ-specific expression of Arabidopsis genome during development', *Plant Physiol.* **138**(1), 80–91.

Mao, C. X. & Lindsay, B. G. (2007), 'Estimating the number of classes', *Ann. Statist.* **35**(2), 917–930.

Mao, C. X. & You, N. (2009), 'On Comparison of Mixture Models for Closed Population Capture-Recapture Studies', *Biometrics* **65**(2), 547–553.

Morris, J. S., Baggerly, K. A. & Coombes, K. R. (2003), 'Bayesian shrinkage estimation of the relative abundance of mrna transcripts using sage', *Biometrics* **59**(3), 476–486.

Otis, D. L., Burnham, K. P., White, G. C. & Anderson, D. R. (1978), *Statistical Inference From Capture Data on Closed Animal Populations*, Wildlife Monographs.

Pledger, S., Pollock, K. H. & Norris, J. L. (2003), 'Open capture-recapture models with heterogeneity. I. Cormack-Jolly-Seber model', *Biometrics* **59**(4), 786–794.

Ramsey, F. & Severns, P. (2010), 'Persistence models for mark-recapture', *Environmental and Ecological Statistics* **17**, 97–109.

Rissanen, J. (1983), 'A universal prior for integers and estimation by minimum description length', *Ann. Statist.* **11**(2), 416–431.

Rocchetti, I., Bunge, J. & Böhning, D. (2011), 'Population size estimation based upon ratios of recapture probabilities', *Ann. Appl. Stat.* **5**(2B), 1512–1533.

Royle, J. A., Dorazio, R. M. & Link, W. A. (2007), 'Analysis of Multinomial Models With Unknown Index Using Data Augmentation', *Journal of Computational and Graphical Statistics* **16**(1), 67–85.

Sanathanan, L. (1972), 'Estimating the size of a multinomial population', *The Annals of Mathematical Statistics* **43**, 142–152.

Seber, G. A. F. & Whale, J. F. (1970), 'The removal method for two and three samples (Corr: V27 p1104)', *Biometrics* **26**, 393–400.

Sepulveda, N., Paulino, C. D. & Carneiro, J. (2010), 'Estimation of T-cell repertoire diversity and clonal size distribution by Poisson abundance models', *J. Immunol. Methods* **353**(1-2), 124–137.

Simar, L. (1976), 'Maximum likelihood estimation of a compound Poisson process', *Ann. Statist.* **4**(6), 1200–1209.

Skibinsky, M. (1986), ' Principal representations and canonical moment sequences for distributions on an interval', *J. Math. Anal. Appl.* **120**(1), 95–118.

Tancredi, A. & Liseo, B. (2011), 'A hierarchical bayesian approach to record linkage and population size problems.', *The Annals of Applied Statistics* **5**(2B), 1553–1585.

Tardella, L. (2002), 'A new Bayesian method for nonparametric capture-recapture models in presence of hterogeneity', *Biometrika* **89**(4), 807–817.

Tardella, L. & Farcomeni, A. (2008), 'On identifiability of population size from capture-recapture data with heterogeneity by the use of marginal likelihood approaches'.

Thygesen, H. & Zwinderman, A. (2006), 'Modeling sage data with a truncated gamma-poisson model', *BMC Bioinformatics* **7**(1), 157.

Wang, J.-P. (2010), 'Estimating species richness by a Poisson-compound gamma model', *Biometrika* **97**(3), 727–740. With supplementary data available online.

Wang, J.-P. (2011), 'Species: An r package for species richness estimation', *Journal of Statistical Software* **40**(9), 1–15.

Wang, J.-P. Z. & Lindsay, B. G. (2005), 'A penalized nonparametric maximum likelihood approach to species richness estimation', *Journal of the American Statistical Association* **100**(471), 942–959.

Yang, H.-C. & Chao, A. (2005), 'Modeling animals' behavioral response by Markov chain models for capture-recapture experiments', *Biometrics* **61**(4), 1010–1017.

Yip, P. S. F., Xi, L., Chao, A. & Hwang, W.-H. (2000), 'Estimating the population size with a behavioral response in capture-recapture experiment', *Environmental and Ecological Statistics* **7**(4), 405–414.

Zeng, L. & Cook, R. J. (2007), 'Transition Models for Multivariate Longitudinal Binary Data', *Journal of the American Statistical Association* **102**(477), 211–223.

Zhao, F.-Z. (2012), 'Some recursive formulas related to inverse moments of the random variables with binomial-type distributions', *Statistics and Probability Letters* **82**(7), 1290–1296.